STATISTICS IN SCHOOL

STATISTICS
IN
SCHOOL

By W. L. SUMNER

Senior Lecturer in the Department of Education
at University College, Nottingham

OXFORD
BASIL BLACKWELL
1948

To my teacher, Sir Cyril Burt,
Professor of Psychology,
University College, University of London

CONTENTS

CHAPTER I

INTRODUCTION

(THE NATURE OF MENTAL MEASUREMENT)

With numbers all men may contend, their charming systems to defend.
GOETHE

We may look at the science of statistics from two angles. Firstly, it may be regarded as the process of collecting figures which represent such things as amounts of exports, price levels, temperatures and barometric pressures from day to day, examination marks and so on, for which some scale of measurement has been found in a world which becomes progressively more metrical. Secondly, statistics is the study of the means of manipulating and arranging figures, applying mathematical processes and thereafter interpreting the results.

Scientific workers try to use the most effective language for their particular purposes. Clear verbal description is a necessity of course, but the precise language of mathematics is also necessary both to describe and to manipulate the results of observations. Scientists usually feel that they are on firm ground when they can provide a 'measuring stick' in order that they can give quantitative results at the end of their experiments and observations. It must be remembered that these results are completely dependent not only on the accuracy of the observations, but also on the size and accuracy of the 'measuring stick'. There is nothing absolute about their findings; they are merely a matter of comparison with an agreed unit of a scale, which in itself is an arbitrary measurement accepted by a large number of workers as a convenient common standard. In the physical sciences where we begin with measurements of length, which lead to those of area, volume and mass, and the measurement of time, there are considerable difficulties in fixing standards. (We assume, for instance, that time has certain properties of length and direction, and may be thought to have some of the properties of a straight line. Great and bewildering

new discoveries were made by Einstein and others in the fi of physics, when some of the elementary foregone conclusic concerning measurements of length and time were challenged.)

In the study of the 'properties' of the human mind, the probl is much more difficult. The mind is not a thing to be measur and weighed as can the whole physical human body, or even brain. When we talk about the factors of the mind, the abilit or intelligence of man, we have to be careful to avoid the pitfall thinking of these as so many tangible quantities each capa of measurement in terms of length, volume or force and so on. is only fairly recently that the 'faculty' psychology (which v kept alive by educationists long after its natural term of yea has been properly buried. The mind must not be thought of terms of a series of faculties, such as intelligence, memory or v and it would be unfortunate if we were to bury 'faculties' and resurrect 'factors' in their place.

The study of arithmetic should always be sustained by logi thought, but many people tend to accept figures and numb uncritically. It has been said cynically that statistics are the wo form of falsehood. This ought not to be correct, but the posit may always be safeguarded by a critical examination of the thi or ideas which underlie them. A simple example of this v suffice. Some years ago some statistics were used in an unscru lous endeavour to show that insulin therapy was useless in ca of diabetes. It appeared that more people had died each y from this disease since the introduction of insulin than befor had been discovered. Moreover, the figures were correct they stood! A little thought will show that the figures had b used to sustain a false argument. Diagnosis of the complaint l improved and thus diabetes had later been given as a cause death, whereas before, the condition was attributed to heart fail pneumonia or internal inflammation. Also, terminating as a ca of death may be, from the statistical point of view, what re matters is whether insulin has extended useful lives, perhaps u fairly advanced age, even though death eventually takes place it must for everybody from one cause or another). The conclu is that insulin is useful.

Investigations in the physical sciences are on the whole easier than those on the measurement of human and social factors. In the physical sciences we are usually able to isolate the property which we wish to measure and to insulate it, so to speak, from disturbing external influences. Different physical properties do not usually cause mutual perturbations which worry the physicist. In any case he can allow for them accurately. He is not usually worried about the barometric pressure of the room, the colour, the magnetic and electrical properties of a piece of metal when he is measuring its specific heat. Moreover, he is able to use units which can be measured in a linear way and about which there is universal agreement.

The matter is not so simple for the psychologist, educationist and even the biologist, for they find it difficult, or even impossible to proceed from cause to effect.[1] The quantities which we think we have isolated and measured today have changed by the morrow. When we believe that we have isolated a physical system in the living body or a 'factor' in the mind, the integration of function and the working unity of the whole have to be taken into account even when we hope that we are studying some specific small 'part'. The twofold aspects of mental activity, the *cognitive* or intellectual and the *orectic* or striving and emotional have to be thought of as being distinct when we try to measure various manifestations of either of them. It does not need much experience and thought to see that there are enormous difficulties in isolating their factors. It is one of the triumphs of modern experimental psychology and statistical analysis, that in a large measure we have been able to clear away misconceptions concerning the so-called 'factors of the mind' and to substitute ideas which are based on scientific principles. Although we cannot always resist the temptation to 'name' certain well-marked aspects of mental activity, we must avoid the temptation to think of these aspects as concrete quantities even if we discover a scale by which they can be estimated on a quantative basis. We shall meet this exceedingly important consideration again.

All mathematical problems which try to provide information

[1] In the last analysis the physicist does too.

concerning the world external to the investigator can be thou of in three stages:

(1) The collection of data, taking care that we have the pr 'measuring rod' for the job in hand and that we know hov use it.

(2) By mathematical processes, the manipulation of the fig of the data, and eventually the arrival at a numerical result.

(3) The interpretation of the result in relation to the orig data. We apply the result to give us further information o predict possible future happenings.

At length we may go from generalizations to tentative 'la Unfortunately, the second step is the only one which has l stressed in schools in the past. Really, it is but a link in a n important and lengthy chain of reasoning.

To make this matter clear let us take as an example a prob from psychological research. Suppose we wish to find whe there is any general measure of agreement (correlation) betw ability in classical studies and general intelligence. In the stage of our investigation we have to evolve a suitable examina in classics for each age group, which will ensure that every has a fair chance and that there are sufficient questions examinees to avoid errors of sampling. The examination pa should be suitable for ready marking on a scale which i keeping with certain statistical requirements. The measuren of intelligence is not such an easy matter. Nevertheless, with enlarging on the considerable difficulties which beset a task wl many people imagine to be relatively simple, we will assume th set of marks in classics and a score in an intelligence test give the same large number of pupils have been obtained.

The second stage is the mathematical process whereb coefficient of correlation between the marks in classics and scores in intelligence tests is obtained.

The last stage is to ask whether this coefficient is signific how many times larger is it than the probable error, what is meaning and value of this correlation, what relationship has i

other possible correlations, and to what conclusions and further investigations of educational significance if any, will it lead.

Although we have used the term 'yardstick' loosely in dealing with mental characteristics, it must be noted that there is a great difference between mental measurements and those of tangible and physical quantities. For instance, a length of seven feet is equivalent to the sum of the lengths of seven separate feet, but a similar consideration does not apply to the type of numerical abstraction which is obtained in the measurement of human abilities or sensory discrimination. Mental measurements have to be made by indirect means and are further complicated by the fact that the very things which are measured are ill-defined and that psychologists may even differ as regards the definitions of the factors which it is proposed to measure. The measurement of so-called 'general intelligence' is a case in point. All psychological measurement involves sampling and it is necessary to take steps to ensure that the sample is fully representative of the group, and secondly that it is large enough to reduce errors of sampling to small proportions Moreover, it is necessary to know what are the possible errors which may mar an estimate made with samples of particular sizes. In addition to errors which are due to sampling there are other difficulties. We must know the degree of validity of a test as a measure of a particular characteristic. It has been claimed that tests have been evolved which are a 'measure of pure *g*' (Spearman's general factor). On investigation, it is found that such tests are 'loaded' (or saturated) with *g* to little more than 70% of their whole variance. Again, a test should have self-consistency or reliability. If it is divided into two parts by taking the odd and even numbered questions separately, there should be a high degree of agreement between the results scored in each half of the test. Although consistency in a test is essential to its validity it is not, of course, sufficient to determine the latter. We shall deal with these matters in a later chapter.

Finally, in educational measurement there is always the possibility of irrelevant factors disturbing the estimation of particular characteristics. Hitherto, most mental measurements have dealt with the cognitive or intellective factors of mental activity, and

it is difficult to separate these from conative or emotional distu ing elements. Finally, even the simplest individual is a rich complex integration of mind and body which is fluctuating f day to day, or even from moment to moment. The physic brass weight is not sensibly different today from what it yesterday, but the human body-mind can never be the same, it may have changed considerably.

CHAPTER II

DISTRIBUTIONS AND DISPERSIONS OF SCORES

If we measure the heights of a large number of boys of the same age, we find that they are distributed in a definite way. We can imagine the boys lined up against a long wall starting with the smallest boy and making each successive boy slightly higher than the last, in going from left to right. The line joining the tops of their heads will be a curve with a shape which would be an elongation of the following:

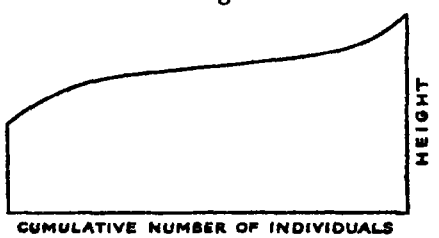


Fig 1. An ogive or cumulative frequency curve The curve can also be drawn with the number of cases given vertically (ordinates) and the marks or other measures given horizontally (abscissae)

It is known as an OGIVE (because of a similar curve which appeared in classical architecture). We could obtain the same curve by picking a thousand ears of wheat from a field (or a large number of peapods of the same crop) and arranging each of them vertically in a horizontal row, starting with the smallest and finishing with the longest. In biology and psychology we can think of many measurements of a similar kind made on a large number of things of the same type, which would give an ogive if plotted in this way. We shall meet this curve again when we are dealing with percentiles. It is sometimes known as a *cumulative frequency curve*. It is often more useful to find the *frequency* or the number of cases occurring in each range whether of height, weight, marks, intelligence, quotient, etc. An easy way is to plot a HISTOGRAM.

Consider the following distribution of marks in which each step is one of 10 marks.

Marks	*No. of Pupils*
0- 10	3
10- 20	12
20- 30	21
30- 40	28
40- 50	35
50- 60	37
60- 70	29
70- 80	17
80- 90	10
90-100	5

The height (and therefore the area) of each column gives a measure of the number of pupils whose marks lie between the figures at the foot of the column. The whole area of the rectangular columns gives the total number of pupils. Here a word of warning is necessary, and it is wise to keep in mind the scales which are used for the marks along the horizontal axis and for the frequencies which are vertical measurements. The value of a unit area on the graph will serve as a guide. The histogram is sometimes spoken of as a *Column Diagram.*

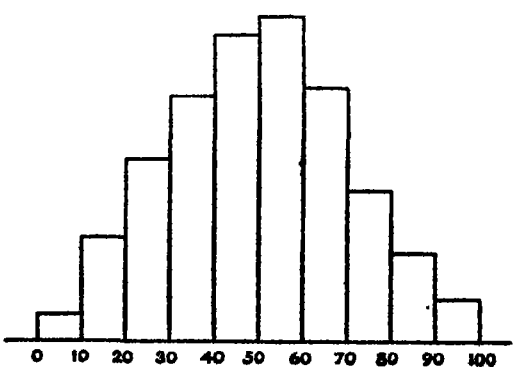


Fig. 2. Histogram.

Suppose we now consider the mid-points of the top of each column to be joined by straight lines and completed at each end by further straight lines joined to the horizontal line as shown in the diagram. We then have a FREQUENCY POLYGON. The frequency polygon does not give quite the exact picture of the data which is yielded by the histogram, especially when the number of cases is small, but frequency polygons may be superimposed and compared and this is a useful property.

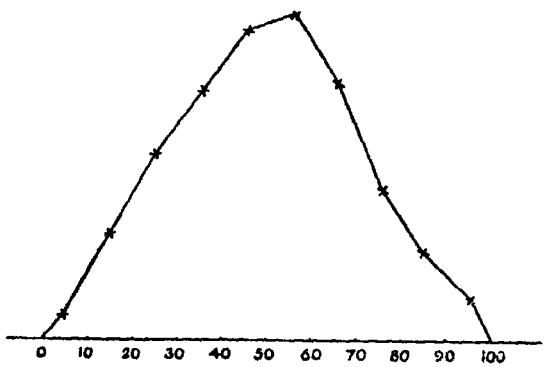


Fig 3 Frequency Polygon

It will readily be appreciated that if we take a large number of cases which show distribution in a regular manner, the frequency polygon will take such a shape that it suggests a 'smoothness' which would tend to a curve if the intervals of marks became smaller as the numbers of cases became larger.

We now come to a most important case of frequency distribution. This is represented by the *curve of normal distribution* or what was formerly called the *curve of error* or the *probability curve.*

Suppose we measure the heights of 10,000 adult Englishmen and plot a histogram showing the number in each half-inch range from (say) 58 inches to 77 inches. (It is possible that we may even

have to extend the range to include men smaller than 4 feet 10 inches, and those taller than 6 feet 5 inches.) If we can now join the mid-points of the tops of the columns and then smooth the frequency polygon to make a curve we should get a shape like the following:

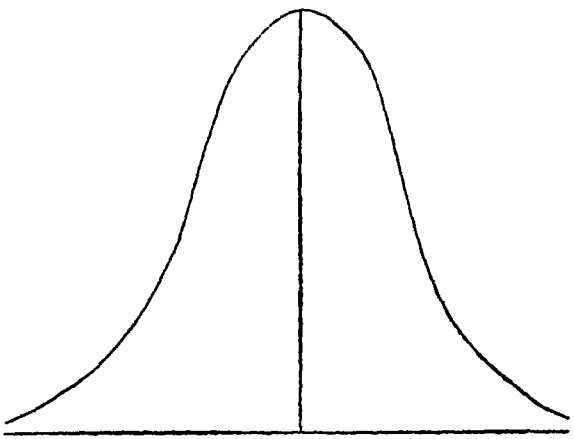


Fig 4.

This distribution is of the utmost importance in science. We shall refer to it as the curve of normal distribution. It used to be called the curve of error because it showed astronomers the distributions of the errors in their readings about the correct value, or again, in gunnery it gave the frequencies of the missiles in respect of their distances from the target after the range had been found.[1] The curve is also known as the probability curve for reasons which will be apparent in a later section of this book. If a curve is not symmetrical about a line drawn through its highest point it is said to be SKEWED and is known as a SKEW CURVE.

[1] For the properties of this important curve see Chapter V and the appendix.

Below is a *positively skewed curve* and the greatest frequency occurs before we come to the middle 'score':

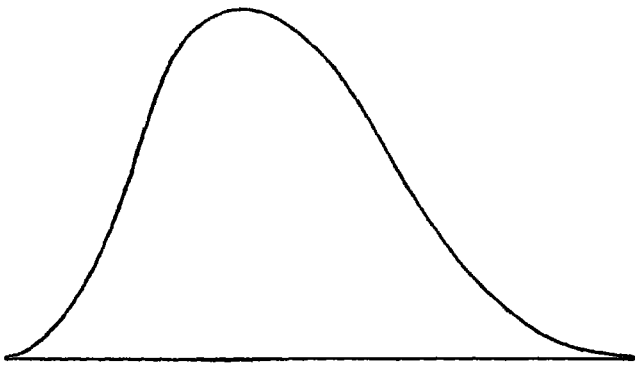


Fig. 5. Positively skewed curve.

and this is a *negatively skewed curve* and the greatest frequency occurs after the middle score.

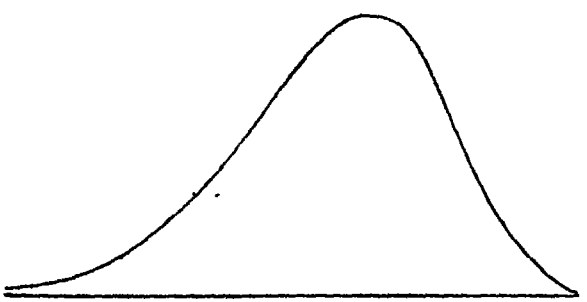


Fig. 6. Negatively skewed curve.

We shall see how the skewing of curves of examination marks and test scores affects the value of the investigation, when we come to apply these matters to the problems of marking.

A curve like the following is known as a *bimodal curve* because it contains two humps, modes or most 'popular' scores. We might obtain such a curve if we gave an intelligence test to a large number of children which consisted of two groups whose abilities were sharply divided.

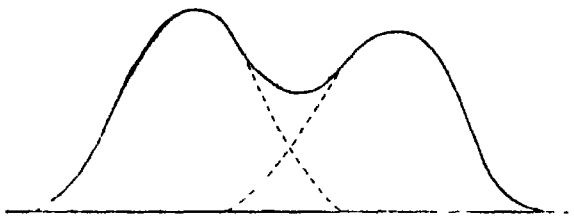


Fig. 7 A bimodal curve

It will be observed that the curve of normal distribution is symmetrical about a vertical line drawn through its highest point. If instead of the heights of a large number of Englishmen, the curve were made to represent the scores of a large number of children in an examination, this line would be a measure of the maximum number of children in any of the mark groups. In the case of the symmetrical curve we see that (*a*) the mark which was scored by the greatest number of children was the average mark of 50%, (*b*) the middle child in an order of merit list scored the average mark. This is obvious as the area enclosed by the curve to the left of the central straight line is equal to the area enclosed by the curve to the right of this line.

It will be noticed that in this and other curves there is a central tendency. The average value (score, mark, height, etc.) is called the MEAN. The value of the middle case (e.g. the mark of the pupil who is half way down an order-of-merit list or rank) is called the MEDIAN. The score, mark, height, etc. which relates to the largest number of individuals is called the MODE. (This is also the meaning of the word in ordinary life.)

Example: The following is a list of marks obtained by school-children in a geography test. Find the mean or average.

Pupil	%
A	45
B	70
C	21
D	32
E	51
F	68
G	48
H	39
I	17
J	84
K	64
L	60
M	44
N	92
O	15
P	31
16	781

Divide 16)781(48·8

Average 48·8%

Add each column down and check by adding up: tick the column total when agreement is reached.

If the marks are represented by x

The Mean $M = \frac{\Sigma x}{N}$ where Σ (sigma) is the sum of (the scores) and N is the number of pupils.

An easier way of calculating an average (especially where there is no great spread of the measures) is to guess the mean and then adjust it by summing the differences of each measure from this mean and dividing by the number of measures, e.g.

Find the mean of the following marks:

Pupil	*Marks*	*Guessed Average*	*Difference* +	*Difference* −
A	61	50	11	
B	40	50		10
C	52	50	2	
D	37	50		13
E	71	50	21	
F	47	50		3
G	54	50	4	
H	32	50		18
I	73	50	23	
J	45	50		5
K	64	50	14	
L	38	50		12
M	41	50		9
N	50	50		
O	46	50		4
P	53	50	3	
			78	74

16 pupils. $+4$

$$\therefore \text{Mean} = 50 + \tfrac{4}{16} = 50\tfrac{1}{4}$$

This method may be expressed as follows:

$M = A + \frac{\Sigma D}{N}$ where A is the guessed or arbitrary mean and D is the sum of the differences (deviations) of each measure from this mean.

Median

The median is the mid-point in a distribution and the number of cases above it is equal to the number below it. It is easy to find the mid-point of a distribution which has an odd number of cases, e.g. 3. 4. 5. 5. 7. 8. 8. 9. 10 for clearly 7 is the value of the median which is the fifth case.

If N is the number of cases and is odd, the median is the $\frac{N+1}{2}$th case. In a distribution with an even number of cases, we must take the mean value of the scores just above and just below the centre point.[1]

e.g. in 3. 4. 5. 5. 7. 8. 8. 9 the median falls between 5 and 7 and can reasonably be given the value 6.

From this we can extend our division of the distribution into quartiles and percentiles. In the following distribution:

2. 2. 4. 5. 6. 7. 8. 9. 10. 10. 12. 13. 14. 14. 16.

it is easy to see that 5. 9. 13 respectively are the values which lie $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$ of the way along the distribution.

The measure representing the first quartile Q_1 is the $\frac{N+1}{4}$th.

The measure representing the second quartile (median) is the $\frac{N+1}{2}$th.

The measure representing the third quartile Q_3 is the $\frac{3(N+1)}{4}$th.

When the number of measures increased by one is not exactly divisible by 4 the same formulae hold: in the case of a large number of cases it will usually suffice to give the value at each quartile point as that of the nearest case. When we have a smaller number of measures an estimate of the values can be made by simple interpolation.

We may extend the division of the distribution into percentiles (100 divisions), or deciles (10 divisions).

The xth percentile is the measure which is $\frac{x(N+1)}{100}$ from the beginning or lower end of the distribution. It is often convenient to plot percentile scores on a piece of graph paper on which a frequency curve or histogram is also drawn.

If we know the marks at the 1st, 10th, 25th, 50th, 75th, 90th

[1] Median is not quite the same thing as 'mid-score' as the median is strictly a point and the mid-score will have a discrete value.

and 99th percentiles, we have an excellent idea of the distribution and by plotting a graph we can find a score corresponding to a percentile, and a percentile (which gives us an idea of order and merit or rank in the distribution) corresponding to a given score.

In a normal distribution a difference in percentile rank corresponds to a greater difference in scores at the beginnings and ends (the tails) of the distribution than in the middle. In fact as regards mark equivalents the 1st, 6th, 22nd, 50th, 78th, 94th and 99th are about equally spaced. We cannot therefore take the averages of a pupil's percentile levels in various subjects in the same way that we can combine his scores.

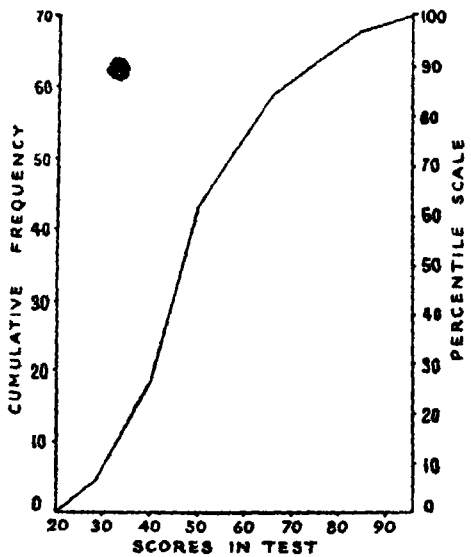


Fig 8. Percentile scale for class of 70 Here the scores are plotted horizontally and the percentile levels and their equivalent cumulative frequencies are plotted vertically. A point on the graph will give the score at any percentile level, or the total number of people who have not reached a certain mark

Finding percentiles when data are given in tabulated form

The results of examinations and tests are often given in tabulated form and sometimes the statistical treatment of sets of marks is easier if they are put into group frequencies.

Consider the following scores in an intelligence test. They are given as the frequencies (the number of persons tested) falling into each score range of 5 marks:

Test Scores	*Frequency*	*Cumulative Frequency*
135-139	0	0
130-134	5	5
125-129	8	13
120-124	9	22
115-119	12	34
110-114	18	52
105-109	25	77
100-104	18	95
95- 99	20	115
90- 94	13	128
85- 89	6	134
80- 84	7	141
75- 79	2	143

Total number N = 143.

The majority of percentile levels will fall inside one of the classes or score ranges. In the above example with an awkward number such as N = 143 all of them will fall within a class.

We can find the percentile (rank) corresponding to a given score from the following formula:

$$x_p = L + \frac{C}{f}\left(\frac{PN}{100} - S\right)$$

where P = percentile, x_p the value of the test score or other measure falling at this percentile level.

L is the lower limit of the class in which x_p lies.

S is the sum of all the frequencies (the number of persons tested) up to but not including this class.

f the frequency within this class.

N the total of all the frequencies.

C the size of this class.

Example: Find the 77th percentile score if

$$P = 77 \quad N = 143 \quad L = 115 \quad f = 12 \quad C = 5 \quad S = 109$$

$$x_p = 115 + \frac{5}{12}\left(\frac{77 \times 143}{100} - 109\right)$$

$$= 115 + \frac{5}{12}\left(111{\cdot}1 - 109\right)$$

$$= 115 + \frac{5}{12}\left(2{\cdot}1\right)$$

$$= 115{\cdot}9 \text{ (116 approximately)}$$

Percentiles also offer a useful way of comparing sets of marks no matter what are the scales of marking.

It is obvious that there is some advantage in giving a student's score in terms of a percentile for then the middle of the rank would always be the 50th percentile. The unfamiliarity of this method to the layman or the uninitiated would probably lead to errors in its interpretation. Although percentiles give a ready means of comparing distributions they must not be used for combining them. Obviously percentile units are much closer to one another near the middle of a distribution than they are at each end.

The mean, median and mode are various ways of regarding the central tendency in a distribution but it is also necessary to have a measure of the spread or dispersion of the set of marks or other measures. In order to secure a proper arrangement of a number of pupils in order of merit it is obviously necessary that the marks should not be bunched together at any point but should be properly distributed. Again, when we come to consider the problems of error in estimating psychological 'factors', it is necessary to know how the errors are distributed. These are two of the many instances of the use of methods of estimating dispersion in mental measurement.

Interquartile Range

The quartile deviation is widely used. If the scores are arranged in rank or order of merit, the difference in score between the first and third quartile points is known as the *interquartile range*. We arrange the scores in order of merit, find the score which is a quarter of the way along the distribution and that which is three-quarters of the way along the distribution and subtract the scores. Dividing by two gives the *quartile deviation* Q (or the semi-interquartile range):

$$Q = \frac{Q_3 - Q_1}{2}$$

It will be observed that Q_2 is the score of the mid-point in the order of merit list. It is therefore the median score.

It will also be seen that as half the scores on one side lie between the median and the first quartile point, and between the median and the third quartile point on the other, the interquartile range gives a measure of dispersion. If the median score can be taken as a particular measure and the other scores in the distribution as differing from it (either above or below) as deviations or errors, the interquartile range will contain half the deviations and therefore the semi-interquartile distance can be regarded as the *probable error*.[1]

Mean Deviation or Average Deviation from the Mean. (*Mean Variation*)

The deviations (differences) of the scores from the mean or average are all regarded as positive and added together. This sum is divided by the number of individuals or cases.

$$\text{Mean deviation} \quad \text{M.D.} = \frac{\Sigma d}{N}$$

Standard Deviation

This measure of dispersion or spread is of great importance and is that which is usually of the most value for mathematical treatment and for the calculation of correlation coefficients. In finding

[1] The range from the 10th to the 90th percentiles called D by some writers is a useful measure of dispersion.

the Mean Deviation above, we regarded each of the deviations as having a positive sign, which was not actually true. If each of the deviations is squared this difficulty is overcome. Moreover, the squaring of each deviation will tend to give due weight to any comparatively large deviation. It also remains to be said that the use of Standard Deviation is in keeping with the mathematical properties of the curve of normal distribution.

To find standard deviation each deviation is calculated and squared. The column of squares is summed and this sum is divided by the number of cases and finally the square root is taken. S.D. is 'root-mean-square' and is usually represented by the small Greek letter sigma σ

$$\sigma = \sqrt{\frac{\Sigma d^2}{N}}$$

Sometimes when we are comparing sets of scores it is necessary to add a subscript to sigma, thus σ_1 or σ_2 to indicate to which group of marks the standard deviation refers. Readers who are not familiar with mathematical notation need not be worried about the sign Σ which is the large Greek sigma S and means 'the sum of —'.

Students should consider the following four methods of computing the standard deviation, and choose that which appears to be the easiest and most labour-saving in view of the given data.

1. The direct method. The mean (or average) is found, the deviation of each score or mark from the mean is calculated, these are squared, added and the square root is found.

In all these methods of calculating the standard deviation a set of tables of squares and square roots such as Barlow's, logarithms and/or a simple slide-rule will be useful. It is hardly ever necessary to give the answer correct to more than two places of decimals and usually one will suffice.[1]

[1] A word of warning ought to be given concerning the finding of square roots. A rough mental estimate will always give the clue to the particular square which is required and where the decimal point should be placed.

To square a number by logarithms, double the log of the number and find the antilog. To find the square root halve the logarithm of the number and then find the antilog. See the appendix for the use of the slide-rule for this and other purposes.

Example (for the sake of simplicity a very 'short' list of scores is taken):

Mark	D	D^2	
8	$8-6=$ 2	4	Mean $= \frac{30}{N} = 6$
7	$7-6=$ 1	1	
4	$4-6$ 2	4	$\frac{\Sigma D^2}{N} = \frac{34}{5} = 6{\cdot}8$
9	$9-6=$ 3	9	
2	$2-6=$ 4	16	$\sigma = \sqrt{\frac{\Sigma D^2}{N}} = \sqrt{6{\cdot}8} = 2{\cdot}6$
Total 30		34	

2. Usually the mean does not turn out to be a whole number and the squares of the deviations contain decimal fractions which cause considerable labour. In this case we guess a mean *which is a whole number* and then apply a correction. A quick mental calculation will suffice to supply the arbitrary mean.

	Mark	D	D	D^2	
$N = 6$	10	$10-6$	4	16	$\frac{\Sigma D^2}{N} = \frac{35}{6} = 5{\cdot}83$
	3	$3-6$	3	9	
	7	$7-6$	1	1	
	8	$8-6$	2	4	
	5	$5-6$	-1	1	
	4	$4-6$	-2	4	
				35	

Guessed mean A = 6

True mean $M = \frac{37}{6} = 6{\cdot}17$

The formula for S.D. in this case $\sigma = \sqrt{\frac{\Sigma D^2}{N} - (M-A)^2}$

$\therefore \sigma = \sqrt{5{\cdot}83 - (6{\cdot}17-6)^2} = \sqrt{5{\cdot}83 - {\cdot}03}$

$= \sqrt{5{\cdot}8} = 2{\cdot}41$

3. When there are only a few numbers to be considered and all the scores or marks are whole numbers, it will suffice to call the arbitrary 'mean' zero. Thus, the deviations (D) will be the original marks (x) and the formula then becomes

$$\sigma = \sqrt{\frac{\Sigma x^2}{N} - M^2}$$

Mark x	x^2
10	100
3	9
7	49
8	64
5	25
4	16

$$M = \frac{37}{6} = 6.17 \qquad \Sigma x^2 = 263$$

$$\sigma = \sqrt{\frac{263}{6} - 6.17^2}$$
$$= \sqrt{43.83 - 38.03}$$
$$= \sqrt{5.8}$$
$$= 2.41$$

4. The mean can be calculated at the same time as the standard deviation by using a modification of the formula on page 21 which now becomes

$$\sigma = \sqrt{\frac{\Sigma D^2}{N} - \left(\frac{\Sigma D}{N}\right)^2}$$

which is obvious when we remember that

$$\text{True Mean} = \frac{\Sigma D}{N} + \text{A (Arbitrary Mean)}$$

and D is the deviation from the guessed or arbitrary mean.

Calculation of the Standard Deviation when the measures are given in grouped frequencies

Even with the use of tables, slide-rules and calculating machines there is considerable labour in calculating the S.D. of a large number of measures. This may often be simplified by putting them into frequency groups. Or it may happen that the measures are originally given in this form.

The formula then becomes:

$$\sigma = \sqrt{\frac{\Sigma fD^2}{N} - \left(\frac{\Sigma fD}{N}\right)^2}$$

in terms of the size of the interval (or extent of each group).

If we wish to express the formula in the same units as the measure (i.e. in score form) the formula is

$$\sigma = \sqrt{\frac{\Sigma fD^2}{N} - \left(\frac{\Sigma fD}{N}\right)^2} \times i$$

where i is the size of the interval of each group.

When a calculating machine is used the easiest form of this expression is

$$\sigma = \frac{i}{N}\sqrt{N\Sigma fD^2 - (\Sigma fD)^2}$$

In each case all the scores in the interval are taken to have a value equal to that given by the mid-point of the interval. D is the deviation of each measure from an arbitrary mean and f the frequency, i.e. the number of measures in each class or interval.

Example: In the following table the marks are given in the first columns, the mid-points of the intervals in the next and then the frequency in each interval. Find the S.D.

Marks	*Mid-Point of Interval*	f	D	fD	fD^2
91-100	95·5	1	+ 4	4	16
81- 90	85·5	2	+ 3	6	18
71- 80	75·5	3	+ 2	6	12
61- 70	65·5	6	+ 1	6	6
51- 60	55·5	11	0	0	0
41- 50	45·5	12	− 1	− 12	12
31- 40	35·5	10	− 2	− 20	40
21- 30	25·5	6	− 3	− 18	54
11- 20	15·5	3	− 4	− 12	48
1- 10	5·5	1	− 5	− 5	25
		$N = 55$	$\Sigma fD = -45$		$\Sigma fD^2 = 231$

$$\left(\frac{\Sigma fD}{N}\right)^2 = \left(\frac{-45}{55}\right)^2 = .67$$

$$\frac{\Sigma fD^2}{N} = \frac{231}{55} = 4{\cdot}20$$

$$\text{S.D.} = 10\sqrt{\frac{\Sigma fD^2}{N} - \left(\frac{\Sigma fD}{N}\right)^2} = 10\sqrt{4{\cdot}20 - .67} = 10\sqrt{3{\cdot}53}$$

$$= 10 \times 1{\cdot}88$$

$$= 18{\cdot}8$$

Sheppard's Correction for Grouped Data

When the measures are grouped into a frequency distribution the S.D. calculated by the method above is somewhat larger than it would have been had the measures been dealt with separately. It can easily be seen that when the deviations are squared, those that lie beyond the mid-point will add relatively more to the sum than those that lie on the 'smaller' side.[1] In the case of a normal distribution Sheppard has shown that in terms of interval units the σ^2 should be diminished by $\frac{1}{12}$ of its value. Thus the corrected S.D. will be given by $(\sqrt{\sigma^2 - \frac{1}{12}}) \times i$ where σ is the crude S.D. found from the grouped frequencies. This is equivalent to

$$\text{corrected S.D.} = \left\{\sqrt{\frac{\Sigma fD^2}{N} - \left(\frac{\Sigma fD}{N}\right)^2 - \frac{1}{12}}\right\} \times i$$

or

$$\frac{i}{N}\sqrt{N\Sigma fD^2 - (\Sigma fD)^2 - \frac{N^2}{12}}$$

As we shall see later when we are studying normal distribution the standard deviation is a most important measure of dispersion. For instance, if we assume normal distribution and know the value of the mean (which in this case will also be equal to median and mode values) we can calculate in terms of the standard deviation the y value (number of cases) for any x value (score or

[1] The matter is further complicated by the fact that each interval in the diagram has a trapezoidal shape.

marks). If we assume, for instance, that intelligence quotient I.Q. is distributed normally[1] and we know the standard deviation and can assume a mean of 100, we can at once calculate the percentage of population possessing particular intelligence quotients, or with I.Q.s between one level and another. This will be understood by a consideration of the properties of the curve dealt with in Chapter V.

Standardized and Normalized Scores

If the scores in a test are represented as measures below or above their average, and they are then divided by their standard deviation, they are represented by z_1, z_2, etc. and are said to be *standard* (or z) *scores*. Approximately two-thirds of the scores will lie between 1 and -1. If the scores can be taken to be distributed normally each set of scores can be regarded as equivalent and comparable. Standard scores can be regarded as deviations from the mean which have been adjusted so that the standard deviation is unity. (It is possible that to call the average mark 0 and to make all marks below it negative, may have a bad psychological value, but in the statistical handling of scores it is often the most convenient way.) Sometimes the scores are *normalized* by dividing their differences from the mean by $\sigma\sqrt{N}$, that is, by the product of the standard deviation and the square root of the number of persons. Standardized scores can be converted to normalized scores by dividing by the root of the number of persons. In the case of *normalized scores* it will be seen that the sum of the scores is unity,[2] and as we shall see later the sum of their products is the correlation coefficient.

The *variance* of a set of scores is the square of the standard deviation. Where a set of scores has been standardized the variance will clearly be unity. We shall use this again when we meet factorial and variance analysis.

It may be useful to return to the question of percentiles and to think of them in terms of standard scores.

[1] There is evidence that this is not quite true.
[2] See Appendix VI.

Assuming a normal distribution:

Percentile Level	*Mark*	*Standard Score. Deviation from mean* (50) ÷ S.D. 10
99	73	+ 2·3
90	63	+ 1·3
75	57	+ 7
50	50	0
25	43	− ·7
10	37	− 1·3
1	27	− 2·3

The limits of the distribution are taken to be + 3 S.D to − 3 S.D.

(In the area under the normal curve (see Chapter V) only ·135% of the measures lie outside this range.)

For psychological reasons the mean might be taken as 60 instead of 50, all the marks then being raised by 10. This does not affect the distribution.

Intelligence tests differ with respect to both their mean and their standard deviation. Scores can only be compared by standardization. In the Moray House Tests the mean is taken as 100 and the S.D. 15.

Percentile	*Score*	*Standard Score*[1]
99 (approx.)	135	+ 2·3σ
95	125	+ 1·7σ
90	120	+ 1·3σ
84	115	+ 1·0σ
75	110	+ ·7σ
50	100	0
25	90	− ·7σ
16	85	− 1·0σ
10	80	− 1·3σ
5	75	− 1·7σ
1 (approx.)	65	− 2·3σ

[1] Some writers do not differentiate between standard and standardized scores, but this need not cause the reader any confusion. A standard score really means a score given as a deviation from the mean with the standard deviation as unit, *i.e.* deviation divided by standard deviation. Standardized scores mean those that have been adjusted to an agreed mean and standard deviation. Before such adjustment the scores are called raw scores.

It will be observed that the scores with standard deviation from the mean fall at the 16th and 84th percentile levels.

Sometimes it is necessary to convert these sigma or z scores to a scale with a given mean and a given standard deviation. Such an operation would also obviate the necessity of using negative scores and those with decimal fractions. Such scores were called t scores by McCall in *How to Measure in Education*. All that is necessary is to multiply each z score with the given S.D. and add to or subtract from the given mean.

Measure of Skewness

If a distribution is symmetrical, its median, mode and mean are at the same point. If a distribution has a positive skew, that is, if it has a long tail stretching towards the high scores, its median will be less than its mean and its mode will usually lie between these.

$$\text{Skewness} \quad \text{Sk} = \frac{\text{mean} - \text{mode}}{\text{standard deviation}}$$

$$= \frac{\text{M} - \text{Mo}}{\sigma}$$

$$\text{or} \qquad \text{Sk} = \frac{3(\text{M} - \text{Md})}{\sigma}$$

where Md is the median.

[A less useful measure of skewness is given by

$$\text{Sk} = \beta_1 = \frac{(\Sigma x^3)^2}{N^2\sigma^6}$$

where the x's are deviations from the mean, and N is the number of measures in the distribution.]

The shape of a symmetrical distribution is measured by its *kurtosis* or flatness β_2

$$\beta_2 = \frac{\Sigma x^4}{N\sigma^4}$$

For normal distribution $\beta_2 = 3$.]

$$\text{Mode} = \text{mean} - \frac{\text{mean} - \text{median}}{C}$$

For many curves and for moderate degrees of skewness $C = \frac{1}{3}$
Thus, to compute the mode from the mean and the median

$$\begin{aligned} \text{Mode} &= M - 3(M - Md) \\ &= 3Md - 2M \end{aligned}$$

(which could have been obtained by equating the first two expressions given above for Sk.)

Coefficient of Variability

The relation which the probable deviation bears to the mean score is of interest as it gives a measure of the variability. We have already seen that the semi-interquartile range Q is equal to the probable deviation (P.E.).

Thus the variability is $\frac{Q}{M}$

If this is expressed as a percentage it is called the *coefficient of variability*.

$$V = \frac{100Q}{M}$$

This is quite independent of the measures used, whether they are marks or the weights of human beings. In general, if V is greater than $\frac{1}{4}$ or 25% the dispersion is regarded as being rather large and the results should be used with great caution.[1]

[1] V is also used for Variance, and its two uses should not be confused

CHAPTER III

CORRELATION AND REGRESSION

If we consider the marks in science and mathematics gained by the members of a class, we should feel justified in expecting that there may be some relation between them. We should hardly anticipate that the top boy in science would also be the top boy in mathematics and that all the boys would have the same orders in both subjects until we came to the unfortunate boy who was at the bottom of the list in science and also in mathematics. If this curious relationship between the mark lists in these subjects did exist, with its exact correspondence of one order to the other, we should say that the marks were *perfectly correlated positively*. If the orders of the marks in both subjects were reversed, the top boy in one subject was the bottom boy in the other, the second boy in the science list was the last but one in the mathematics list, and so on (this is unthinkable, of course!), we should say that here was a case of *perfect negative correlation*. If the marks in science bore no relation at all to those in mathematics we should say that there was *no correlation*. In practice we should expect to find some *positive* connection between marks in these two subjects, but it would be partial or imperfect correlation. This type of correlation is most important when we consider examination marks, and the scores in psychological and other tests; and exact mathematical methods for dealing with it are of the utmost importance in many educational and psychological researches. The correlation coefficient is almost as important to the psychological tester as is the balance to the chemist. As we shall see in a later chapter, many extraordinary assertions were made by educationists and psychologists in the past, and continue even today, because statements concerning human abilities or 'intelligence' had not been subjected to rigorous analysis in which the use of correlation coefficients is invaluable. Nevertheless, other techniques are sometimes more valuable, but a clear idea of correlation is none the less of prime importance.

We can obtain a useful graphical idea of the degree of correlation between sets of numbers by plotting a *scatter diagram* or *scattergram*. Suppose we plot the scores in two subjects or tests of a number of individuals, giving a point on a two-dimensional graph to each individual. The co-ordinates of each point (x,y) are measures of the scores in each subject. Suppose further that the scores have been standardized by calling the mean (average) of each set zero, and then dividing each deviation from zero by the standard deviation of the set.

If there were no correlation between the x and y values (the scores in each test) the points representing the individuals would be distributed in a haphazard manner over the graph paper, that is to say, there would be a fairly even density of points on the graph paper, provided that we had taken results from a sufficiently large number of individuals. If there existed some degree of correlation between the x and y scores, we should find that the points tended to bunch together and were more dense in a certain

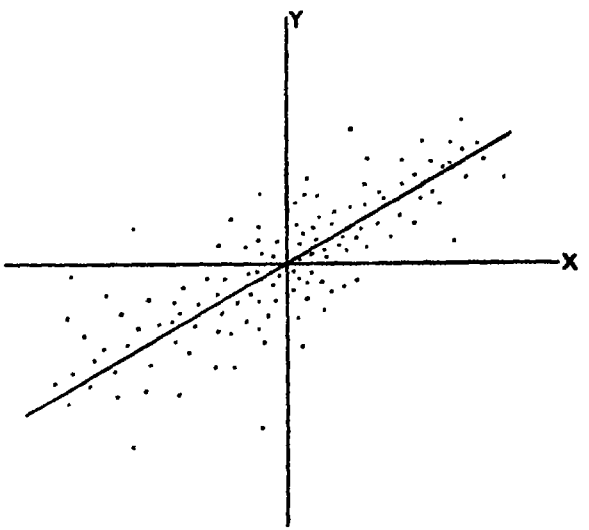


Fig. 9.

direction than in others. When we plot points in this manner we are said to have made a *scatter diagram* or *scattergram.*

Where correlation is present we can find a line which best fits the distribution of the points which we have plotted. Few, if any, points will lie on it, but the line will go through the cluster of points so that there is a 'balance' of points on each side of it. It would then be the *line of best fit*, and would be known as the *line of regression.*[1] (Although this term is not particularly apt for psychological work, it is invariably used. It is a biometric term used by Galton to show that the average heights of offspring tend to 'regress back towards the mean of the race'.)

Suppose that the correlation is a perfect positive one. The points would be bunched together in the first and third quadrants and the line of best fit would make an angle of 45° with the positive x axis. If, on the other hand, there was perfect negative correlation, the points would be bunched together in the second and fourth quadrants, and the line of regression would be at right angles to that representing perfect positive correlation. In education and psychology we usually find that correlation, if present, is *partial positive correlation.* Thus we shall find the lines of regression in the first and third quadrants (or if we are dealing with 'raw' or uncreated scores upwards from 0 in the first quadrant).

The slope of the regression line, that is its $\frac{y}{x}$ value, or the tangent of the angle which it makes with the x axis, is equal to r, the correlation coefficient.[2]

In the case of perfect positive correlation, writing x_1 and x_2 as deviations from their means and σ_1, σ_2 as the respective standard deviations,

$$\frac{\frac{x_1}{\sigma_1}}{\frac{x_2}{\sigma_2}} = 1 \ (= \tan 45^\circ)$$

[1] The sum of the squares of the distances of the points from the line should be a minimum.

[2] See Appendix I.

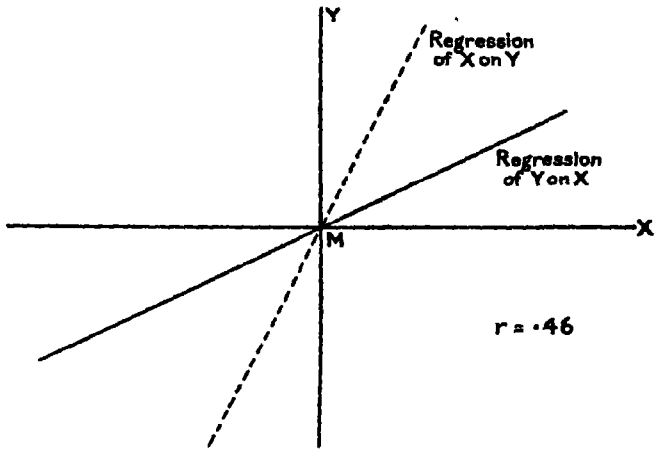


Fig. 10.

When there is no correlation we should have to regard the regression line as being horizontal, with a slope of 0.

In general the slope of the regression line of x_1 on x_2 is given by

$$\frac{x_1}{\sigma_1} \Big/ \frac{x_2}{\sigma_2}$$

Thus the coefficient of correlation r is

$$\frac{x_1}{\sigma_1} \Big/ \frac{x_2}{\sigma_2}$$

or $x_1 = rx_2 \dfrac{\sigma_1}{\sigma_2}$. This is the equation of the regression line.

The regression line of x_2 on x_1 makes the same angle with the vertical axis as the regression line of x_1 on x_2 does with the horizontal axis. The equation of this line of regression (x_2 on x_1) is

$$x_2 = rx_1 \frac{\sigma_2}{\sigma_1}$$

Before leaving the subject of regression it may be useful to note that regressions do not seem to obey the ordinary algebraic rules, for instance, the regression of x on y may be written $x = ry$ and that of y on x will be $y = rx$. Thus the regressions occur in pairs

$$x = ry$$
$$y = rx$$

The following diagrams will help to explain the phenomenon of regression.

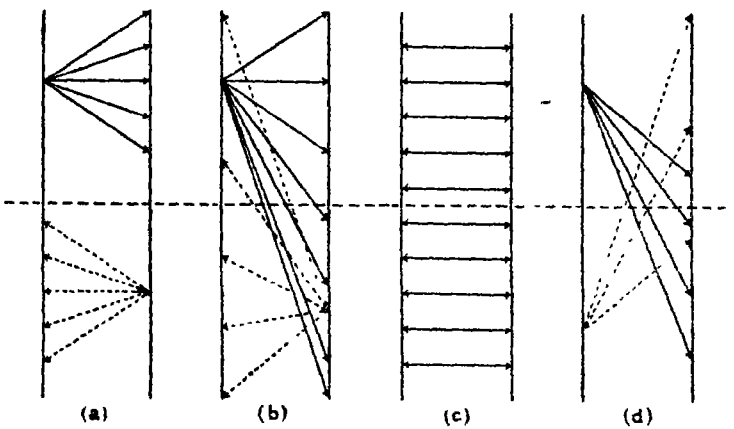


Fig. 11

In each case the vertical lines represent two sets of scores. In (*b*) there is no correlation, and any set of frequencies in one set may be matched with any score throughout the range of the other. In (*a*) where there is some positive correlation the frequency group in one set corresponds with a spread of scores in the second, but for the most part in the same side of the mean. In (*d*) there is negative correlation with the spread tendency to be on the opposite side of the mean. In (*c*) there is perfect correspondence between the scores and correlation is complete *and there is no regression*. Thus it can be seen that from regression equations we can estimate the value of one or other variable for an individual when we know the correlation coefficient between

the two variables. As the correlation coefficient increases, our accuracy of estimation of the one variable will improve as shown in the table and diagram on page 48. It will repay the student to consider the significance of the correlation coefficient, as it appears in regression equations, and its use in the prediction of the amount of regression will give a clearer idea of its nature than that which comes from its calculation from formulae. This will also serve to explain the apparent paradox of the two regression equations such as $y = rx$ and $x = ry$, for just as there is an uncertainty varying with the amount of regression in predicting an x value from a y, so also there exists a similar uncertainty in predicting a y value from an x.

A reference to scatter diagrams will also serve to reveal whether correlation is linear. We shall see that although it is usually safe to assume that it is so, this is not invariably the case, and the line of best fit is then not a straight line.[1] Although the correlation coefficient is a measure of the degree of relationship between two sets of measures, it is not directly proportional to the degree of relationship. For instance, a correlation coefficient of ·7 does not represent twice the degree of relationship given by a correlation of ·35. It is also necessary to interpret the correlation coefficient in relative terms. A correlation coefficient of 9 would not be high in the case of two 'paired' and similar mental tests, whereas in determining the degree of relationship between a physical and a mental characteristic it would be difficult to find a value of r much greater than ·5. It is common to speak of a value of r less than ·3 as low, from ·3 to ·7 as medium, from ·7 to ·9 as high, and above ·9 very high, but without reference to the meaning of the sets of measures which have been correlated, such terms may be entirely misleading.

As we have already noted, the correlation coefficient enables us to predict with a degree of reliability which is known (and should be allowed for) the most likely value of a variable in one set, when that in the other set and the correlation coefficient between the sets are known. The diagram and table on page 48 illustrate this.

[1] See page 56.

The Product-Moment Method of Calculating the Correlation Coefficient (sometimes known as the Bravais-Pearson Method[1])

From the general consideration of the degree of agreement between sets of measures or scores arranged so that the average of each is adjusted to be zero, which we have seen when we were dealing with regression, it is apparent that if there is no measure of agreement between one set of measures and another, the sum of the products of deviations of corresponding scores (carrying appropriate signs) from the mean, will tend to be zero. If there is a tendency for measures above the mean in one set to correspond to measures which are above the mean in the other (taking the mean as zero), and those below in one to correspond with those below in the other, it is obvious that the total of the products of the deviations of each score in one set and the corresponding score in the other will be a positive number. Thus, the product of the deviations will give an idea of the existence of a positive correlation. In the same way, if the positive deviations of one set tend to correspond with negative deviations of the other their product will be a negative number and will give an idea of the negative correlation between them.

The exact formula, known as the PRODUCT-MOMENT or Bravais-Pearson formula for the correlation coefficient, which is written as r is

$$r = \frac{\Sigma x_1 x_2}{N\sigma_1\sigma_2}$$

where x_1 and x_2 are the deviations of the respective scores in each case from the mean of each set, N is the number of cases, e.g. the number of pupils in a class, and σ_1 and σ_2 are the standard deviations of the respective sets of scores. If the scores have already been *standardized* by dividing their deviations from their respective means by the standard deviations, the formula becomes

[1] Bravais, a French statistician of the nineteenth century first used the idea of product-moments, and his work was improved by Galton. Karl Pearson (1857-1925), scientist and statistician, may be regarded as the successor of the latter. The name product-moment refers to the products of the moments (or the weights) of the scores in relation to their deviation from the mean

$$r = \frac{\Sigma z_1 z_2}{N}$$

where z_1 and z_2 are the *standardized* scores, and further, if the scores have been normalized by dividing the standardized scores by $\sqrt{N}$, the correlation coefficient will then become $r = \Sigma s_1 s_2$ where s_1 and s_2 are the *normalized* scores.

Where the correlation coefficient is calculated from the deviations x_1 and x_2, the means will hardly ever be whole numbers, and the exact determination of $\Sigma x_1 x_2$ is apt to be a laborious process. When calculating standard deviations we saw how it was possible to use an arbitrary or guessed mean which was a whole number, and if x_1 is now a deviation from an arbitrary mean the standard deviation

$$\sigma_1 = \sqrt{\frac{\Sigma x_1^2}{N} - \left(\frac{\Sigma x_1}{N}\right)^2}$$

The formula for the correlation coefficient therefore becomes

$$r = \frac{\dfrac{\Sigma x_1 x_2}{N} - \dfrac{\Sigma x_1}{N} \times \dfrac{\Sigma x_2}{N}}{\sqrt{\dfrac{\Sigma x_1^2}{N} - \left(\dfrac{\Sigma x_1}{N}\right)^2} \sqrt{\dfrac{\Sigma x_2^2}{N} - \left(\dfrac{\Sigma x_2}{N}\right)^2}}$$

Example. Use the simple formula for calculating the correlation coefficient from the data given opposite.

$$r = \frac{\Sigma xy}{N \sigma_x \sigma_y}$$

$$= \frac{\Sigma xy}{\sqrt{\Sigma x^2} \sqrt{\Sigma y^2}}$$

$$= \frac{6035}{\sqrt{11,671} \times \sqrt{7616}}$$

$$= \frac{6035}{108{\cdot}0 \times 87{\cdot}3}$$

$$= {\cdot}6401$$

Example: CALCULATION OF CORRELATION COEFFICIENT BY PRODUCT-MOMENT METHOD USING THE SIMPLE FORMULA

Pupil's Number	%*Maths* X	%*Physics* Y	X − *Mean* X x	Y − *Mean* Y y	x^2	y^2	xy
1	38	50	− 5	− 7	25	49	35
2	39	44	− 4	− 13	16	169	52
3	61	62	18	5	324	25	90
4	49	54	6	− 3	36	9	− 18
5	29	50	− 14	− 7	196	49	98
6	51	72	8	15	64	225	120
7	64	70	21	13	441	169	273
8	59	70	16	13	256	169	208
9	29	64	− 14	7	196	49	− 98
10	27	60	− 16	3	256	9	− 48
11	19	74	− 24	17	576	289	408
12	61	60	18	3	324	9	54
13	43	36	0	− 21	0	441	0
14	11	48	− 32	− 9	1,024	81	288
15	42	46	− 1	− 11	1	121	11
16	46	70	3	13	9	169	39
17	72	76	29	19	841	361	551
18	62	42	19	− 15	361	225	− 285
19	33	64	− 10	7	100	49	− 70
20	40	40	− 3	− 17	9	289	51
21	37	62	− 6	5	36	25	− 30
22	39	52	− 4	− 5	16	25	20
23	46	72	3	15	9	225	45
24	71	78	28	21	784	441	588
25	25	28	− 18	− 29	324	841	522
26	19	30	− 24	− 21	576	441	504
27	66	80	23	23	529	529	529
28	73	78	30	21	900	441	630
29	52	60	9	3	81	9	27
30	28	46	− 15	− 11	225	121	165
31	53	64	10	7	100	49	70
32	20	64	− 23	7	529	49	− 161
33	56	48	13	− 9	169	81	− 117
34	24	38	− 19	− 19	361	361	361
35	57	64	14	7	196	49	98
36	11	46	− 32	− 11	1,024	121	352
37	39	56	− 4	− 1	16	1	4
38	60	72	17	15	289	225	255
39	29	56	− 14	− 1	196	1	14
40	27	32	− 16	− 25	256	625	400
	1,707	2,284		Totals	11,671	7,616	6,035
	Mean X = 43 (rounded)	**Mean Y = 57 (rounded)**					

It will be observed that even with only 40 measures in each set and the slight inaccuracy introduced by taking whole numbers for the mean considerable labour is involved; tables of squares and square roots, logarithms and a slide-rule may be used to reduce the labour of computation. A calculating machine which will add, multiply, square (and if possible divide) is of great use where much of this work is done.

X=	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	Fy
Y=+5									1				1
+4				1			1						2
+3					1			2		1			4
+2		1				1			1				3
+1					4	1	2		1				8
0					2	3	1						6
-1				2	1		2	1					6
-2						2							2
-3				2									2
-4	1			1				1					3
-5			2										2
-6		1											1
Fx	1	2	2	6	8	7	6	4	3	1	0	0	40 Total (N)

Fig 12.

To avoid this type of calculation it is better to draw a scatter diagram of the data to be correlated and proceed as follows. (Often the data will have been given in grouped frequencies at the start and therefore the grouping of the measures in the form of a scatter diagram on squared paper is the obvious next step).

Here the measures have been grouped into 12 rows and 12 columns. These numbers need not have been equal but 11 or 12

should be regarded as a minimum, otherwise the grouping will be too coarse and as the S.D.s calculated by this method will be too large the correlation coefficient will tend to be too small. The two sets of measures to be correlated are denoted by X and Y. Convenient arbitrary means are chosen and the deviations of each group of measures are given above and below the respective means taken as O. The figures in the cells give the numbers (frequencies) of the cases with the corresponding X and Y values.

It should be seen that the separate totals of the X and Y values each come to N (the total number of cases).

Stage 1. *To calculate the Standard Deviations*

We use the Standard method of grouped frequencies as given in Chapter II.

Deviations		*Frequencies*		*Frequencies* × *Deviations*		*Frequencies* × (*Deviations*)^2	
x	y	f_x	f_y	$f_x \cdot x$	$f_y \cdot y$	$f_x \cdot x^2$	$f_y \cdot y^2$
+ 6		0		0		0	
+ 5		0	1	0	5	0	25
+ 4		1	2	4	8	16	32
+ 3		3	4	9	12	27	36
+ 2		4	3	8	6	16	12
+ 1		6	8	6	8	6	8
				—	—		
0		7	6	27	39	0	0
				—	—		
− 1		8	6	− 8	− 6	8	6
− 2		6	2	− 12	− 4	24	8
− 3		2	2	− 6	− 6	18	18
− 4		2	3	− 8	− 12	32	48
− 5		1	2	− 5	− 10	25	50
− 6			1		− 6		36
		—	—	—	—	—	—
	N =	40	40	− 39	− 44	172	279

$\Sigma f_x \cdot x = -12 \quad \Sigma f_y \cdot y = -5$

$$\frac{\Sigma f_v \cdot x}{N} = \frac{-12}{40} = -.300 \qquad \left(\frac{\Sigma f_x \cdot x}{N}\right)^2 = .09$$

$$\frac{\Sigma f_v \cdot y}{N} = \frac{-5}{40} = -.125 \qquad \left(\frac{\Sigma f_y \cdot y}{N}\right)^2 = .016$$

(We shall need the value $\frac{\Sigma f_x \cdot x}{N} \times \frac{\Sigma f_y \cdot y}{N}$ for the numerator of the correlation formula and in this case it is very small: $-.3 \times -.125 = .037$.)

$$\frac{\Sigma f_v \cdot x^2}{N} = \frac{172}{40} = 4.30 \qquad \frac{\Sigma f_y \cdot y^2}{N} = \frac{279}{40} = 6.97.$$

$$\sigma_x = \sqrt{\frac{\Sigma f_x \cdot x^2}{N} - \left(\frac{\Sigma f_x \cdot x}{N}\right)^2}$$

$$= \sqrt{4.30 - .09} = \sqrt{4.21} = 2.05$$

Similarly $\sigma_y = \sqrt{6.97 - .016} = \sqrt{6.954} = 2.64$

Stage 2. To find the sum of the total x and y products

The frequency (number of cases) in each cell must be multiplied by the product of its x and y values. This can be done by considering each possible product, and finding the total frequencies of the cells with each value. It is obvious that any cell with a zero value for x *or* y will contribute nothing to the total. The cells may be crossed out in pencil as they are dealt with. The total frequencies should come to N.

A table of three columns may be constructed to give respectively the possible xy products (those which are not represented by actual cases need not be written down), the frequencies, and the product $f \times x \times y$.

xy	f	fxy	
0	10	0	
+ 1	3	3	
+ 2	2	4	
+ 3	1	3	
+ 4	1	4	
+ 6	5	30	
+ 8	1	8	
+ 12	1	12	
+ 15	3	45	
+ 20	1	20	
+ 24	1	24	153
− 1	6	− 6	
− 2	1	− 2	
− 3	1	− 3	
− 8	3	− 24	−35
	N = 40	Σfxy = 118	

$$\frac{\Sigma fxy}{N} = \frac{118}{40} = 2{\cdot}95$$

After the correction has been subtracted the numerator is $2{\cdot}95 - {\cdot}037 = 2{\cdot}913$

$$r = \frac{2{\cdot}913}{\sigma_x \times \sigma_y} = \frac{2{\cdot}913}{2{\cdot}05 \times 2{\cdot}64} = {\cdot}54.$$

Another method which is sometimes simpler than the above, is to apply the formula

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_d^2}{2\,\sigma_x \times \sigma_y}$$

where σ_x and σ_y are the S.D.s as before and σ_d is a third S.D. calculated as follows:

By means of a ruler or straight edge inclined at an angle of 45° to the horizontal the number of measures falling in diagonals taken from the top left-hand corner to the bottom right-hand corner are considered. (Each diagonal will be at right angles to the line drawn from the top left-hand corner to the bottom right-hand corner. The first diagonal containing any measures will be that drawn from $y = +1$ to $x = -2$ and this will contain two measures, the next from $y = 0$ to $x = 0$ contains no measures.)

The measures will read as follows from the table on page 38:

2 0 1 1 6 8 6 5 9 1 0 0 1 making the correct total of 40.

By choosing an arbitrary mean the S.D. is calculated as before and this will supply the value σ_d for the formula, which is then worked.

Rank Correlation

The product-moment method of finding the correlation coefficient is undoubtedly the best way for use in scientific investigations but when the number of cases to be considered is less than 30 the *method of ranks* is just as reliable, and in some cases is even more so. The ranks (or orders of merit) in the two sets of marks or test scores are written against the names of the pupils (It is usually convenient to write the names in order of merit in one subject and in a column to the right to add the correct order in the other subject with which we seek correlation.) The difference in rank is written in the next column and in a fourth this difference is squared. This column which contains only square positive numbers is then totalled. The difference of rank is called d, each difference is squared (d^2) and these squares are summed Σd^2.

If we consider N pupils (or cases) it is easy to prove that if N is not too small, the sum of the differences of ranks squared which would result from pure chance[1] or probability would be $\frac{N(N^2-1)}{6}$

or $\frac{N(N-1)(N+1)}{6}$

[1] See Appendix IV.

[As N is a whole number, notice that $\frac{N(N^2-1)}{6}$ is also a whole number]

The fraction of disagreement between two sets of orders of merit or ranks could therefore be expressed as follows:

$$\frac{\text{Sum of the actual difference squared}}{\text{Sum of the differences squared which might be expected by chance}}$$

or $\frac{\Sigma d^2}{\frac{1}{6}N(N^2-1)}$.

If this is a measure of the amount of disagreement of the ranks, the measure of the agreement or correlation may be written as

$$1 - \frac{\Sigma d^2}{\frac{1}{6}N(N^2-1)}$$

[by subtracting from unity]

This correlation coefficient using ranks is written as ρ (the Greek letter rho).

It is related to r, the correlation coefficient obtained by the product-moment or line of regression methods, by the formula:

$$r = 2 \sin \frac{\pi}{6} \rho$$

Usually this transformation is hardly worth while. By using ordinary tables of sines the method is as follows:[1] Multiply ρ by 30°. Look up the sine of the resulting angle and double it. This gives r. This relation between ρ and r is only true on the average of many occasions.

[1] The angle π (radians) = 180°. $\frac{\pi}{6}$ = 30°.

Example: CALCULATION OF CORRELATION COEFFICIENT BY RANKS METHOD

Name	*Rank in French*	*Rank in History*	d^2
Ashley	28½	26½	4
Ascough	25	24	1
Beaumont	2½	1	2¼
Clifton	9½	2	56¼
Champkins	28½	29	¼
Evans	19½	38	342¼
Foster	31½	39	56¼
Gill	38	6	1,024
Gasper	13½	7½	36
Gray I	28½	11½	289
Gray II	38	19½	342¼
Green	1	4	9
Goodman	38	31½	56¼
Harrison	9½	5	20¼
Hawley	33½	15	342¼
Hill	22½	31½	81
Jackson	36	29	49
Lymn	5	13	64
Marriot	16½	19½	9
MacEwan	38	37	1
Norman I	31½	35	12¼
Norman II	5	21	256
Nelson	11½	33	462¼
Newham I	28½	15	182¼
Newham II	25	22	9
Peak	16½	29	156¼
Powdril	25	26½	2¼
Pickersgill	21	24	9
Pillatt	13½	36	506¼
Rivers	16½	17½	1
Robinson	19½	40	421¼
Shaw	5	9	16
Shrewsbury	16½	17½	1
Stafford	33½	34	¼
Thornton	11½	11½	—
Walker	7½	3	20¼
Wilcox	7½	15	56¼
Wright	35	24	12
Warkinson	22½	7½	225
Wardle	2½	10	56¼
			5,190¼

$$\rho = 1 - \frac{\Sigma d^2}{\frac{1}{6} N (N^2 - 1)}$$

$$= 1 - \frac{5190\frac{1}{2}}{\frac{1}{6} \times 40 (1599)}$$

$$= 1 - \frac{6}{40 \times 1599} \times \frac{20761}{4}$$

$$= 1 - \cdot 486.$$

Rank Correlation = ·51

Spearman's Footrule

A 'rough and ready' way of finding whether there is any agreement between two sets of results in Spearman's 'Footrule'. It is not intended to yield precise mathematical results but it often gives a quick method of finding whether any correlation exists, and it may be used to prepare the way for more exact investigations.

Only the *gains* in rank are noted, and the *losses* which must in the long run be equal to the gains are neglected. The *gains* are added: Σg.

The coefficient of correlation[1] $R = 1 - \frac{\Sigma g.}{\frac{1}{6} (N^2 - 1)}$

It is rarely necessary to correct the 'Footrule' coefficient R into r the true correlation coefficient but this can be done by using the formula

$$r = 2 \cos \frac{\pi}{3} (1 - R) - 1$$

Here is an example using the data given overleaf:

$$R = 1 - \frac{\Sigma g}{\frac{1}{6} (n^2 - 1)}$$

$$= 1 - \frac{6 (187)}{1599}$$

$$= \cdot 3$$

[1] In this case 'agreement' might be a better word.

Example: SHOWING THE USE OF SPEARMAN'S FOOTRULE.

No. in Class	*Maths.*	*Science*	*Rank in Maths.*	*Rank in Science*	*g.*
1	38	50	25	26½	1½
2	34	23	33	33	10
3	61	50	7½	26½	19
4	49	54	16	24	8
5	29	62	29	17½	—
6	51	72	15	7	—
7	64	70	5	10	5
8	59	70	10	10	—
9	29	64	29	14	—
10	27	60	32½	20	—
11	19	74	37½	5	—
12	61	60	7½	20	12½
13	43	36	19	37½	18½
14	11	48	39½	28½	—
15	42	46	20	31	11
16	46	70	17½	10	—
17	72	40	2	35	33
18	62	42	6	34	28
19	33	62	27	17½	—
20	40	76	21	4	—
21	37	64	26	14	—
22	39	52	23	25	2
23	46	72	17½	7	—
24	71	78	3	2½	—
25	25	28	34	40	6
26	19	36	37½	37½	—
27	66	80	4	1	—
28	73	78	1	2½	1½
29	52	60	14	20	6
30	28	46	31	31	—
31	53	64	13	14	1
32	20	64	36	14	—
33	56	48	12	28½	16½
34	24	38	35	36	1
35	57	64	11	4	—
36	11	46	39½	31	—
37	39	56	23	22½	—
38	60	72	9	7	—
39	29	56	29	22½	—
40	27	32	32½	39	6½
					g. Total = 187

A little consideration of the nature of regression lines will give us a clearer idea of the meaning of correlation than will come from an uncritical acceptance and use of the product-moment formula. It is sometimes thought that a correlation coefficient gives an exact measure in terms of a fraction or percentage of the agreement between two scores. It is indeed true that a correlation coefficient will give us a clue to the common elements which are contained in the scores. As we have seen by drawing lines of regression in scattergrams a correlation coefficient gives us an idea of the reduction of error in predicting scores in one test from those in another.

It is easy by using the formula for probable error to construct a table or draw a graph to show the percentage reduction in error in making this forecast. This is known as the *forecasting efficiency* of the correlation coefficient.

The regression equations are valuable because we can calculate the most probable values of x_2 from x_1 and those of x_1 from x_2. There is likely to be a considerable scatter on both sides of the estimated values of x_1 or x_2 as can be seen by considering an actual scatter diagram.

The Probable Error of the estimated $x_1 = .6745\,\sigma_1\sqrt{1-r^2}$

The Probable Error of the estimated $x_2 = .6745\,\sigma_2\sqrt{1-r_2}$

It can be seen that when $r = 1$ that is when there is perfect correlation $\sqrt{1-r^2} = 0$ and thus there will be no error in finding x_1 from x_2 or x_2 from x_1.

As r decreases the probable error of the estimation becomes greater.

$\sqrt{1-r^2}$ is called the *coefficient of alienation* (Kelley). and is useful in that it gives us an idea of how high r should be for satisfactory prediction.

When $r = .1$ the prediction is only $\frac{1}{2}\%$ (.005) better than pure chance. With r of .8 we are only 40% better than pure chance and with $r = .95$ only 69% better off!

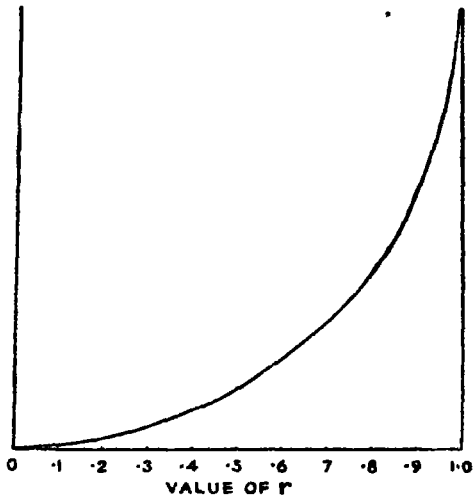


Fig 13 The ordinates (vertical distances) give the forecasting efficiency for various values of the correlation coefficient

Correlation coefficient r	*Forecasting efficiency* %
·00	·0
·10	·5
·20	2·0
·30	4·6
·40	8·4
·50	13·4
·60	20·0
·70	28·6
·80	40·0
·90	56·4
·95	68·8
1·00	100

It can be seen that unless r has a value of at least ·8 the forecasting efficiency will not be above 40% ($\frac{4}{10}$ths) and therefore it will be of little value. With a correlation coefficient of ·3 the forecasting efficiency is less than 5% or a twentieth.

The Correlation of Three Variables

Sometimes we have three sets of correlation coefficients by considering three sets of variables, or attainments taken in three pairs. It may be necessary to find the correlation between any two of the variables supposing that the third were kept constant. Such a case would be to find the correlation between school attainment and estimations of character with intelligence kept constant.

The partial correlation formula is as follows

$$r_{12.3} = \frac{r_{12} - r_{13} \times r_{23}}{\sqrt{(1 - r_{13}^2)\,(1 - r_{23}^2)}}$$

r_{12}, r_{13} and r_{23} are the correlation coefficients of the scores 1, 2 and 3 taken in pairs. $r_{12.3}$ is the correlation coefficient of scores 1 and 2 with 3 kept constant.

As a further example we may consider correlation of age, height and weight. Let us call them x years, y inches and z lb. respectively. We can correlate them in pairs and find r_{xy}, r_{xz} and r_{yz} but each of these correlations is affected by the third variable.

The formula enables us to calculate the correlation between any two, say x and y, left uninfluenced by the third.

In this case the correlation coefficient $r_{xy.z}$.

$$r_{xy.z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$

For convenience of reference we give the standard error now but this will be more fully explained in a later chapter.

$$\text{Standard error} = \frac{1 - r^2}{\sqrt{N}}$$

where r is the particular correlation coefficient which is required.

Tetrachoric Correlation

TETRACHORIC CORRELATION means a method of correlation using four groups (as the Greek name implies). In these methods we have data limited to the number of cases or the proportion of cases in each of two categories in each set.

Suppose we have a number of pupils who are given tests in science and mathematics. We can divide them into four groups.

a = Number above average in both science and mathematics.
b = Number above average in science but below average in mathematics.
c = Number below average in science and above average in mathematics.
d = Number below average in both science and mathematics.

	Science	
Mathematics	a	b
	c	d

Pearson's coefficient is

$$\rho = \text{cosine}\left(\frac{\sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}\right)\pi$$

The value of the expression within the bracket is calculated. This

is multiplied by 180° and the cosine of the resultant 'angle' found from the tables.

It will be seen that the total number of cases (e.g. the number of pupils) $= a + b + c + d$ if we can disregard pupils who are exactly on the average line.[1]

Example: In an examination taken by 40 candidates 6 were above average in both science and mathematics, 14 were above average in science and below in mathematics, 14 were below average in science and above in mathematics, 6 were below average in science and in mathematics.

	Science	
Mathematics	14	6
	6	14

By the formula

$$\rho = \cos\left(\frac{\sqrt{36}}{\sqrt{196} + \sqrt{36}}\right) 180°$$

$$= \cos \frac{6 \times 180°}{20}$$

$$= \cos 54°$$

$$= \cdot 5878$$

A modification of the above is sometimes useful as it gives a conservative (or even modest) idea of the intensity of association. It is known as the *coefficient of colligation* ω and is due to Yule.

[1] When the divisions (i.e., the dichotomic lines) are at the respective means the formula simplifies itself, to

$$\rho = \sin \frac{(ad - bc)}{N^2} 360° \qquad \text{i.e. } \rho = \sin \frac{2\pi (ad - bc)}{N^2}$$

where N = total number of measures $= a + b + c + d$.

$$\omega = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

Using the same data as above:

$$\omega = \frac{\sqrt{196} - \sqrt{36}}{\sqrt{196} + \sqrt{36}} = \frac{14 - 6}{14 + 6} = \frac{8}{20} = \cdot 4$$

The Method of Unlike Signs due to Sheppard

U = percentage of 'unlike' signs (that is, of cases with one score above and one below average in both tests)

$= b + c$

L = percentage of 'like' signs (that is, the sum of cases with both scores above or below average respectively)

L + U = 100 (as U and L are percentages)

Sheppard's coefficient $s = \cos \frac{U}{L + U} \pi$

$$\therefore s = \cos \frac{180U}{100}$$

$$= \cos 1 \cdot 8 \text{ U}$$

Thus, the *percentage* of unlike signs must be found, multiplied by 1·8 (i.e. $\frac{9}{5}$) and the cosine of this number regarded as an angle in degrees found from tables.

In the example used for Pearson's formula above, the percentage of unlike signs is $\frac{12}{40} \times 100\% = 30\%$

$$s = \cos (1 \cdot 8 \times 30)^\circ = \cos 54^\circ$$
$$= \cdot 5878$$

which is precisely the coefficient which we found above. (This does not always happen but usually there is close agreement.)

Coefficient of Association due to Yule

From our tetrachor table we can measure the intensity of association between two sets of data using Q the coefficient of association

$$Q = \frac{ad - bc}{ad + bc}$$

Using the same data as above

$$Q = \frac{14 \times 14 - 6 \times 6}{14 \times 14 + 6 \times 6}$$

$$= \frac{196 - 36}{196 + 36} = \frac{160}{232} = .69$$

This method produces a generous estimate.

Biserial correlation

Sometimes it is necessary to correlate sets of data when they are given in the form of two mutually exclusive groups in respect of one set and in numerical scores in respect of the other. Such dichotomies in the first set would be given by sex differences, married and unmarried persons, trained and untrained teachers, graduates and non-graduates, children of a particular age group attending school and those of the same age who have left school, etc. The following example taken from a study of a hundred boys and girls, sixteen to eighteen years of age who have left school and another group remaining at school will illustrate this.[1]

The biserial coefficient of correlation is given by

$$r_{xy} = \frac{(M_{x_2} - M_{x_1})\, pq}{z\sigma_y}$$

[1] By Elwood Sones. 'A Study of one hundred boys and girls sixteen to eighteen years of age who have left school and a similar group remaining at school' (according to size of families). The correlation between 'Staying at School' and size of family is only .176

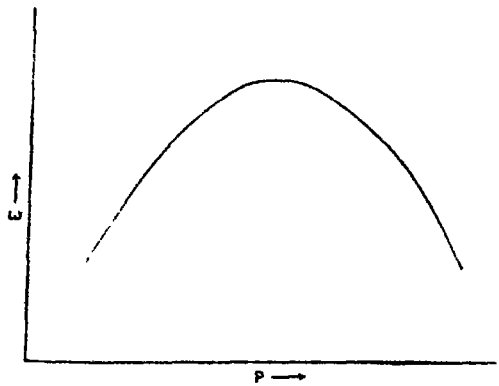


Fig 14. Non-linear correlation

Thus both high and low perseverators would tend to have low character scores, and the highest character scores would be associated with moderate perseveration.

In this case we use a correlation ratio η (eta) which is given by

$$\eta = \sqrt{1 - \frac{\sigma_x^2}{\sigma_y^2}}$$

where σ_x is the standard error of estimate (the standard deviation of one of the sets of measures) and σ_y is the standard deviation of the other.

CHAPTER IV

THE PROBLEM OF ERROR

But to us probability is the very guide to life.
BISHOP BUTLER — *Analogy of Religion*

As the results which we have obtained so far assume that we have been handling very large numbers of cases it is necessary to consider what happens when we make our experiments with smaller samples. It is obvious that the statistical laws which we use will be free from errors to an extent related to the number of cases which we can investigate. A very simple example will suffice to show this. If we toss a penny a sufficiently large number of times, say 100,000, we should expect the ratio of heads to tails to be 1 to 1 with a very tiny possible error in the 1 : 1 ratio. If we toss the coin only 10 times it may happen that we get 3 heads and 7 tails but in the case of 100,000 trials the chances of getting 30,000 heads and 70,000 tails are so exceedingly remote as to have no statistical interest for us. In other words as the number of trials gets larger and larger the ratio of heads to tails approaches nearer and nearer to its true limit.

The problem before us now is to try to find just how reliable are the results of our investigations on various numbers of cases. An ordinary school class may contain no more than 25 or 30 children. Again, when we have to deal with rather lengthy investigations it is necessary to limit the number of cases considered in order that the research can be completed in a reasonable time.

Thus, all the investigations on a metrical basis which we make in psychology and education will have to be qualified by an estimate of the size of the error which is likely to arise, and we shall have to consider its size in relation to the size of other factors concerned, as a correlation coefficient, for instance. In the analysis of variance, that due to error may be compared with the variance due to other causes under consideration.

It is clearly impossible to take such large samples, in normal procedures, to ensure that each sample is a true cross-section of the entire population. Suppose that we have been finding the correlation coefficients between two sets of scores in subjects A and B and that we have been able to continue our investigations with a large number of similar groups of children. We should not find the correlation coefficient to be quite the same in any two groups of children owing to errors of sampling; we should find a central tendency in all the correlation coefficients and it would be apparent that the correlation coefficients would satisfy the normal law of distribution. To find the probable error we should want to know how far from the mean or central value of the correlation coefficient is the line which divides one-half of the coefficients from the rest. If the dispersion were great compared with the value of the correlation coefficient, that is, if the P.E. were more than a small fraction of the correlation coefficient, we should regard the latter as being unreliable.

Investigators trained in the physical sciences tend to reject any results where the correlation coefficient is not more than four times greater than the probable error, but a less rigorous attitude has prevailed in psychological investigations and results which are no greater than three times the probable error are accepted as being significant. Even these should be treated with great caution and the investigation should be continued with further critical exploration of method and data. In writing down a correlation coefficient or other result we should therefore add the value of the probable error.

Probable error is another term for quartile deviation or the semi-interquartile range. Usually, however, the term quartile deviation is only applied to simple measures and probable error is used with derived or secondary measures, as for instance standard deviation or the correlation coefficient. The obvious way of finding the probable error would be to arrange the measures in order or to count them and to take half; but more often the probable error is found from the standard error (or deviation) and the use of the formula P.E. $= \cdot 6745\sigma$ (i.e. $\cdot 6745 \times$ S.D.).

It may be well to examine the meaning of the word *probable*.

If we say that 'It is probable that it will rain tomorrow' we really mean that the chances that it will rain are more than those that it will keep fine, that is to say that the chance is perceptibly greater than a '50-50' chance. The expression 'probable-error' though time-honoured is misleading and really means half the measures on each side of the central point. (A rough approximate is that probable error $= \frac{2}{3} \times$ standard deviation.)[1]

Probable Error of Mean $= \cdot 6745 \frac{\sigma}{\sqrt{N}}$

Probable Error of Standard Deviation $= \cdot 6745 \frac{\sigma}{\sqrt{2N}}$

Probable Error of Correlation Coefficient $r = \cdot 6745 \frac{1 - r^2}{\sqrt{N}}$

The reader must not be misled by the use of the word probable,[2] and the formulae simply give the chances that the mean or other derivatives will lie within a certain distance of the true value. In the case of the mean the chances that it lies between + probable error and − probable error are 1 to 1. The chances that it lies inside the limits become greater as the limits increase: for instance

between − P.E. and + P.E.	the chances are		1 to 1
− 2 P.E. and + 2 P.E.	,,	,,	4·5 to 1
− 3 P.E. and + 3 P.E.	,,	,,	21 to 1
− 4 P.E. and + 4 P.E.	,,	,,	142 to 1
− 5 P.E. and + 5 P.E.	,,	,,	1310 to 1
− 6 P.E. and + 6 P.E.	,,	,,	19,200 to 1

[1] These matters will become clearer when the chapter on the Normal Curve is read. It should be remembered that the relation between standard and probable errors only holds if normal distribution of the errors can be assumed.

[2] The popular treatment of probability in terms of 'odds for' and 'odds against' should be qualified by a more systematic mathematical treatment. Here 'certainty' is denoted by a probability of 1 and an 'impossibility' by a probability of 0. The mathematical probability of an event lies between 0 and 1 and may be expressed as a fraction, decimal fraction or a percentage.

If the probability that an event will happen is given by the fraction $\frac{1}{x}$, (i.e. 1 chance *in* x and not 1 *to* x) the probability against the event happening will be $1 - \frac{1}{x}$ or the fraction $\frac{x-1}{x}$.

The chances that the mean lies outside these limits is given by interchanging the figures in the two right-hand columns.

The chances expressed in terms of standard deviations:

	Frequencies of deviations outside these limits	*Odds against deviations falling outside these limits*
$\pm$ P.E. $= \pm \cdot 6745\sigma$	$2 \times 25\%$	1 to 1
$\pm \sigma$	$2 \times 15{\cdot}9\%$	2 to 1
$\pm 2\sigma$	$2 \times 2{\cdot}28\%$	21 to 1
$\pm 3\sigma$	$2 \times {\cdot}135\%$	370 to 1
$\pm 4\sigma$	$2 \times {\cdot}0032\%$	15,600 to 1

The standard error (or standard deviation) does not tell us how much our result is in error but rather the chances that the result has an error of a particular magnitude.

Summary of the Probable Errors of Correlation Coefficients

r is the correlation coefficient found by the product-moment or line of regression.

$$\text{P.E.} = \cdot 6745 \frac{1 - r^2}{\sqrt{N}}$$

ρ (rho) is the correlation coefficient found by rank method

$$\text{P.E.} = \cdot 706 \frac{1 - \rho^2}{\sqrt{N}}$$

R is the correlation coefficient found by Spearman's footrule or the 'gains in rank' method

$$\text{P.E.} = \frac{\cdot 43}{\sqrt{N}}$$

It will be noticed that in each case the denominator contains $\sqrt{N}$, the square root of the number of cases considered. The consequence of this is that if we quadruple the number of cases (e.g. consider 120 pupils instead of 30) the probable error is reduced by a half, and it will be reduced to a third if the number of cases is multiplied by nine.

PROBABLE ERROR OF THE CORRELATION COEFFICIENT

Number of cases	*r*												
	0 0	0 1	0.2	0 3	0.4	0 5	0 6	0 65	0 7	0 75	0 8	0 85	0 9	0 95
20	1508	.1493	1448	1373	1267	1131	0965	0871	0769	0660	0543	0419	0287	0147
30	.1231	.1219	1182	1121	1035	0924	0788	0711	0628	0539	0444	0342	0234	0120
40	1067	1056	1024	0971	0896	0800	0683	0616	0544	0467	0384	.0296	0203	0104
50	0954	.0944	0915	.0868	0801	0715	0610	0551	0486	0417	0343	.0265	0181	0093
70	.0806	.0798	0774	.0734	0677	0605	0516	0466	0411	0353	0290	.0224	0153	0079
100	.0674	0668	0648	.0614	0567	0506	0432	0391	.0345	0294	0242	0187	0128	0066
150	.0551	.0546	0529	0501	.0463	0413	.0352	0318	.0281	0241	0198	0153	0105	0054
200	.0477	0472	0458	0434	0401	0358	0305	0275	0243	.0209	0172	0133	0091	0047
250	0426	.0421	0409	.0387	0358	0319	0272	0246	0218	0187	.0154	0118	0081	0042
300	.0389	0389	0374	0354	0327	0292	0249	0225	0199	0170	0140	.0108	0074	0038
400	0337	.0334	0324	0307	0283	0253	0216	0195	0172	0148	0122	.0094	0064	0033
500	.0302	0299	.0290	0274	0253	0226	0193	0174	0154	0132	0109	0084	0057	0029
1000	.0213	0211	0205	0194	0179	0160	0137	.0123	0109	0093	.0077	0059	0041	0021

e.g. (*a*) find the P.E. where $r = \cdot 9$, $N = 36$.

$$\text{P.E.} = \frac{\cdot 6745\,(1 - \cdot 9^2)}{\sqrt{36}}$$

$$= \frac{\cdot 6745 \times \cdot 19}{6}$$

$$= \cdot 0213$$

$$r = \cdot 9 \pm \cdot 0213$$

In writing the probable error in this way it must be remembered that the P.E. is given as a probability and not as an actuality.

(*b*) find the P.E. where $r = \cdot 4$, $N = 16$

$$\text{P.E.} = \frac{\cdot 6745\,(1 - \cdot 4^2)}{\sqrt{16}}$$

$$= \frac{\cdot 6745 \times \cdot 84}{4}$$

$$= \cdot 142$$

$$r = \cdot 4 \pm \cdot 142$$

Here the P.E. is more than a third of the correlation coefficient. The latter cannot therefore be considered reliable or even significant.[1] It would have been better in the investigation to have used all possible means to take a greater number of cases than 16.

[1] The nature of the ratio between a coefficient and its standard error or deviation must be carefully considered. The figure which is taken, really means that the chances that the coefficient has no significance are reduced to such an extent that we have reason to believe that there is good evidence of significance. There is no case of conclusive proof. As a figure equal to twice the standard deviation only occurs about once in 22 cases Fisher suggests that this may be regarded as significant. As probable error is about $\frac{2}{3}$ × standard error or deviation, Fisher's suggestion is that 3 × probable error would be a significant quantity.

McCall has suggested a ratio of 2.78 × standard deviation (i.e. about 4.17 probable error), but this is larger than we usually find in psychological and educational experiments even when other considerations lead us to believe that there should be significance and some notable degree of correlation between our figures.

Peters suggests that a figure somewhat less than that of Fisher's may be permitted. He takes the point on the probability curve where it bends to a maximum degree as the distribution thins out to a long tail. This gives a value of 1.73 × S.E. or 2.6 × P.E. and for this he proposes the term *working ratio*.

In each case the student should fortify himself by finding what is the extent of the probability from the tables of the integral of the normal probability curve and it should be kept in mind that probability does not imply certainty.

P.E. of tetrachoric r where the dichotomic lines are at the means

$$\frac{\cdot 6745\ (2\pi\sqrt{1-r^2})}{\sqrt{N}}\left[\frac{(a+d)\ (c+b)}{4N^2}\right]^{\frac{1}{2}}$$

and where true $r = 0$

$$\text{P.E.} = \frac{\cdot 6745\pi}{2\sqrt{N}}$$

The probable error does not give a very good estimate of the reliability of r when N is small and r is large. Accordingly Fisher has suggested that r should be replaced by its hyperbolic arc-tangent $\tanh^{-1} r$ which he calls z^1 and for which he provides tables.

$$\tanh^{-1} r = z^1 = \tfrac{1}{2}\left[\log_e (1+r) - \log_e (1-r)\right]$$

$$= \frac{\cdot 4343}{2}\left[\log_{10} (1+r) - \log_{10} (1-r)\right]$$

Many experimenters would feel that results obtained by investigations with less than 25 cases would be so unreliable as to be of negligible worth and where any rigorous research was undertaken a hundred cases or more should be considered.

Test Reliability and Test Length

If, after a sufficient interval, a test is applied again under similar circumstances there should be a high degree of correlation between the two sets of scores. Moreover, if the test is a good one it should be largely independent of the qualities and skills of those administering it.

If a test is reliable it can only be so if it is thorough and this will depend to a large extent on its length. If tests are supplied in double form so that there are two parallel tests, a re-test with the second set should produce results with a high degree of correlation, that is, upwards of ·9, with the first set. When two similar tests are not supplied, a single test is converted into two by taking the

odd-numbered questions as a shortened first test and the even-numbered as the second test. By shortening the test its reliability is also reduced and therefore it is necessary to have some means of predicting the reliability of a test if it were lengthened.

Suppose r is the correlation coefficient of the results of the two halved tests. Then if R is the correlation coefficient between the complete given test and an imaginary one of similar type

$$R = \frac{2r}{1 + r}$$

In a general case, where a test is imagined to be lengthened n times we may use the *Spearman-Brown prophecy-formula*:

$$R = \frac{nr}{1 + (n - 1)r}$$

(of which the formula for the doubled tests is the simplest case).

We can calculate the reliability or the limits of variation of individual scores when we know the reliability coefficient.

$$\text{Probable error} = .6745\sigma\sqrt{1 - r^2}$$

e.g. if there is a correlation of .95 between intelligence tests and the standard deviation of the intelligence quotients is 15 then

$$\text{P.E. of I.Q.} = .6745 \times 15 \times \sqrt{1 - .95^2}$$
$$= 3.1$$

This means that about half the people taking the second test will have I.Q.s which differ from those which they obtained in the first test by little more than 3 points. By considering the way in which the expression $\sqrt{1 - r^2}$ becomes larger as r becomes smaller the student will see how rapidly the probable error increases as the reliability coefficient r drops below .9. Unfortunately an r of .95 is exceedingly rare. It should be added that the reliability of a test will appear to be lower than it can be taken to be, if it is given to groups which are too homogeneous and therefore do not permit proper sampling both in respect of age and abilities. The

difference in reliability as given by tests with two groups of different 'spread' (i.e. homogeneity or heterogeneity) is given by the formula

$$R = 1 - \frac{\sigma_1^2 (1 - r)}{\sigma_2^2}$$

where R is the reliability to be expected with a group with standard deviation of I.Q. $= \sigma_2$ and r the reliability with a group of S.D. σ_1.

CHAPTER V

THE NORMAL CURVE OF DISTRIBUTION AND ITS USES

Most students are familiar with the well-known bell-shaped curve and we have already noticed it when we were considering the distribution of measures with respect to a central tendency. It is now convenient to consider more carefully the nature of this important curve. For the reader who can deal with simple calculus some of its mathematical properties have been worked out in Appendix III. For the purposes of the present section it will suffice if we examine the shape of the curve and know the meaning of the heights of various lines drawn vertically in it and the significance of areas bounded by the curve and cut off by such lines. The quantitative aspects of such lines and areas will be given in simple tables. The curve is sometimes called the Laplacian or Gaussian curve in honour of Laplace and Gauss who respectively used it in their work on probability. For reasons which will be apparent it is also called the *probability curve* or *curve of error*. One of its most fruitful early uses was to deal with experimental errors in astronomical observations.

A word of warning must be uttered concerning the use of the so-called 'normal' curve. Too often in the past the adjective 'normal' has been misused. The distribution of the velocities of molecules of a gas, or that of the quantitative measures of errors in respect of many physical observations *may under certain conditions* where there are no biasing factors conform to such a curve. Even here the mathematical theory of pure chance in the distribution usually preceded any attempt to check its validity, which has to be assumed without experiment in many cases. In the case of 'mental measurements' the matter is much more difficult. We have no theoretical basis for expecting such distributions, and in fact factors can be imagined which may cause skewing. In an intelligence test scale we are not dealing with the physicist's 'class A' measures such as length, speed and mass. We can obtain

a length of 130 cm. by adding one of 70 cm. to another of 60 cm. We cannot obtain an I.Q. of 130 by adding one of 70 to another of 60. Each I.Q. must be referred separately to an arbitrary scale. It would be foolish to assume that there is a fundamental 'law of normality' which applies to most sets of educational and psychological data. Most of the groups and samples with which we have to deal in psychological research are only defined in a vague and ambiguous manner and the degree of homogeneity in traits other than the one which we are considering is seldom sufficient to eliminate their effect.

It is impossible to talk about the form of a distribution being normal with any meaning unless we specify the type and classification of the individuals concerned.

Certain physical characteristics such as weight show reasonably good normal distribution for individuals of the same sex, race, age and height, but even here the curve is negatively skewed, as in 'normal' times excessive overweight is more common than excessive underweight. The use of the word 'normal' whether it describes the times in which we live, a person's behaviour, or a distribution needs careful consideration. This is not to despise its use in educational research, but the early use of the distribution to deal with errors and deviations from a mean is still the most useful. A curious example of 'circular reasoning' sometimes takes place with respect to intelligence tests. Such tests are usually devised to give a 'normal' distribution of the scores with certain population classes. It is to be expected therefore that when they are applied to the testing of similar population classes the distribution should be normal.[1] The symmetrical bell-shaped curve is useful because it is susceptible to easy mathematical treatment, but here again we must not be ensnared by the attempts which mental testers have made to give numerical assessments of intelligence along a scale of numbers. This scale has none of the properties of a graduated rule or length. The boy with I.Q. 130 is not twice as intelligent as a boy of I.Q. 65. There is in fact

[1] A distribution which does not conform to the 'normal curve' may be quite normal in the usual sense. In educational measurements and calculations the words 'normal distribution' refer to the curve.

hardly any means of comparing these individuals; the first able to benefit by Grammar School teaching and the other practically a moron. The 'man in the street' who said that the first was 'a thousand times as intelligent' as the second would, in spite of exaggeration, have the germ of truth in him.

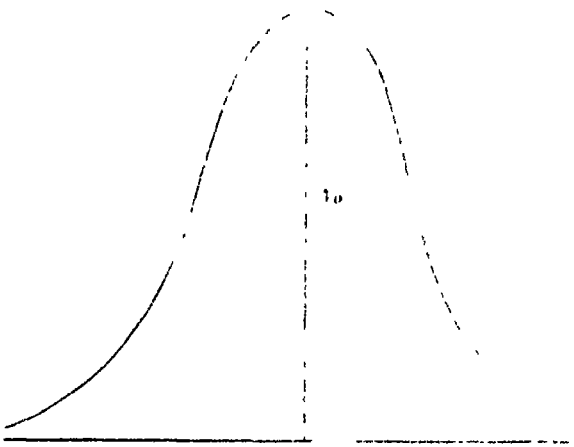


Fig 14a

The mean, mode and median of the curve are equal and are marked y_0 on the central axis of y, about which line the curve is symmetrical. The area of the curve represents the total number of scores or measures which are distributed. By drawing vertical lines we can measure the areas enclosed by the curve which are cut off by them. These represent the numbers of scores which are beyond or within a certain value of the score.

If there is good dispersion of the scores the curve is wide and well-rounded, but if, on the other hand, there is not much dispersion and the scores deviate but little from the mean, the curve is thin, sharp and pointed.

It will be observed that at points on the curve, known as ***points of inflexion***, the convex shape of the top part of the curve gives way to the concavity of the lower part of each side. These points are

at a distance σ (standard deviation) on each side of the central point.

The curve is said to be *asymptotic* to the axis of x (that is the horizontal base line). This means that the curve approaches this line if it is sufficiently extended at both sides It is said to meet the line 'at infinity'. The standard deviation σ is a convenient unit for measuring distances along the x axis. Exceedingly little of the area of the curve remains at distances greater than 3σ on each side of the central line.

It is convenient to reduce all distances along the x axis to *sigma-units* by dividing the x distances by σ.

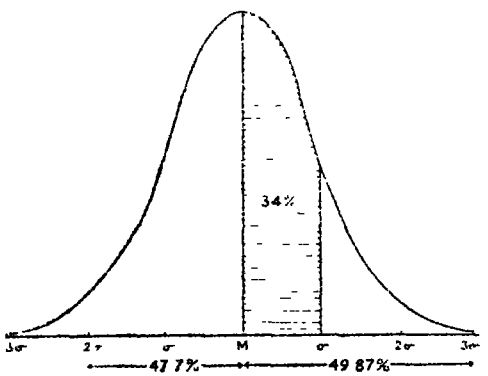


Fig 15

The amount of the area enclosed by the whole curve lying between verticals at distances of σ on each side of the central line is 68·26%.

That enclosed between verticals at distances of 2σ on each side of the central line is 95·44%,

and that enclosed between verticals at distances of 3σ on each side of the central line 99·75%.

The following table gives the proportion (percentage) of the total area under the normal curve between the central line (mean ordinate) and an ordinate (vertical line) at any given distance (in sigmas) from the mean.

TABLE I

PER CENT OF TOTAL AREA UNDER THE NORMAL CURVE BETWEEN MEAN ORDINATE AND ORDINATE AT ANY GIVEN SIGMA-DISTANCE FROM THE MEAN

$\frac{x}{\sigma}$	.00	01	02	03	04	05	06	07	08	09
0 0	00 00	00 40	00 80	01 20	01 60	01 99	02 39	02 79	03 19	03 59
0 1	03 98	04 38	04 78	05 17	05 57	05 96	06 36	06 75	07 14	07 53
0 2	07 93	08 32	08 71	09 10	09 48	09 87	10 26	10 64	11 03	11 41
0 3	11 79	12 17	12 55	12 93	13 31	13 68	14 06	14 43	14 80	15 17
0 4	15 54	15 91	16 28	16 64	17 00	17 36	17 72	18 08	18 44	18 79
0 5	19 15	19 50	19 85	20 19	20 54	20 88	21 23	21 57	21 90	22 24
0 6	22 57	22 91	23 24	23 57	23 89	24 22	24 54	24 86	25 17	25 49
0 7	25 80	26 11	26 42	26 73	27 04	27 34	27 64	27 94	28 23	28 52
0 8	28 81	29 10	29 39	29 67	29 95	30 23	30 51	30 78	31 06	31 33
0 9	31 59	31 86	32 12	32 38	32 64	32 90	33 15	33 40	33 65	33 89
1 0	34 13	34 38	34 61	34 85	35 08	35 31	35 54	35 77	35 99	36 21
1 1	38 43	36 65	36 86	37 08	37 29	37 49	37 70	37 90	38 10	38 30
1 2	38 49	38 69	38 88	39 07	39 25	39 44	39 62	39 80	39 97	40 15
1 3	40 32	40 49	40 66	40 82	40 99	41 15	41 31	41 47	41 62	41 77
1 4	41 92	42 07	42 22	42 36	42 51	42 65	42 79	42 92	43 06	43 19
1 5	43 32	43 45	43 57	43 70	43 83	43 94	44 06	44 18	44 29	44 41
1 6	44 52	44 63	44 74	44 84	44 95	45 05	45 15	45 25	45 35	45 45
1 7	45 54	45 64	45 73	45 82	45 91	45 99	46 08	46 16	46 25	46 33
1 8	46 41	46 49	46 56	46 64	46 71	46 78	46 86	46 93	46 99	47 06
1 9	47 13	47 19	47 26	47 32	47 38	47 44	47 50	47 56	47 61	47 67
2 0	47 72	47 78	47 83	47 88	47 93	47 98	48 03	48 08	48 12	48 17
2 1	48 21	48 26	48 30	48 34	48 38	48 42	48 46	48 50	48 54	48 57
2 2	48 61	48 64	48 68	48 71	48 75	48 78	48 81	48 84	48 87	48 90
2 3	48 93	48 96	48 98	49 01	49 04	49 06	49 09	49 11	49 13	49 16
2 4	49 18	49 20	49 22	49 25	49 27	49 29	49 31	49 32	49 34	49 36
2 5	49 38	49 40	49 41	49 43	49 45	49 46	49 48	49 49	49 51	49 52
2 6	49 53	49 55	49 56	49 57	49 59	49 60	49 61	49 62	49 63	49 64
2 7	49 65	49 66	49 67	49 68	49 69	49 70	49 71	49 72	49 73	49 74
2 8	49 74	49 75	49 76	49 77	49 77	49 78	49 79	49 79	49 80	49 81
2 9	49 81	49 82	49 82	49 83	49 84	49 84	49 85	49 85	49 86	49 86
3 0	49 87									
3 5	49 98									
4 0	49 997									
5 0	49 99997									

The next table gives the ordinates (the vertical heights) under the normal curve at various x distances (in terms of standard deviation) from the mean. The ordinates are given as proportions of the mean ordinate, that is, the greatest height of the curve. Such a table is useful if we desire to find the frequency at a certain point, e.g. the number of cases with a certain score.

TABLE II

ORDINATES UNDER THE NORMAL CURVE AT VARIOUS SIGMA-DISTANCES FROM THE MEAN (ORDINATES EXPRESSED AS PROPORTIONS OF THE MEAN ORDINATE)

$\frac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0 0	1 0000	1 0000	9998	9996	9992	9988	9982	9976	9968	9960
0 1	9950	9940	9928	9916	9903	9888	9873	9857	9839	9821
0 2	9802	9782	9761	9739	9716	9692	9668	9642	9616	9588
0 3	9560	9531	9501	9470	9438	9406	9373	9338	9303	9268
0 4	9231	9194	9156	9117	9077	9037	8996	8954	8912	8869
0 5	8825	8781	8735	8690	8643	8596	8549	8501	8452	8403
0 6	8353	8302	8251	8200	8148	8096	8043	7990	7936	7882
0 7	.7827	7772	7717	7661	7605	7548	7492	7435	7377	.7319
0 8	7262	7203	7145	7086	7027	.6968	6909	6849	6790	6730
0 9	6670	6610	6550	6489	6429	6368	6308	6247	6187	.6126
1 0	6065	.6005	5944	5883	5823	5762	5702	5641	5581	5521
1 1	5461	5401	5341	5281	5222	5162	5103	5044	4985	4926
1 2	4868	4809	4751	4693	4636	4578	4521	4464	4408	.4352
1 3	4296	4240	4185	4129	4075	4020	3966	3912	3859	3806
1 4	3753	3701	3649	3597	3546	3495	3445	3394	.3345	.3295
1 5	.3247	3198	3150	3102	3055	3008	2962	2916	2870	.2825
1 6	2780	2736	2692	2649	2606	2563	2521	2480	2439	2398
1 7	2358	2318	2278	2239	2201	2163	2125	2088	2051	2015
1 8	1979	1944	1909	1874	1840	1806	1773	1740	1708	1676
1 9	1645	1614	1583	1553	1523	1494	1465	1436	1408	1381
2 0	.1353	1327	1300	1274	1248	1223	1198	1174	1150	1126
2 1	1103	1080	1057	1035	1013	0991	0970	0950	0929	0909
2 2	0889	0870	0851	0832	0814	0796	0778	0760	0743	0727
2 3	0710	0694	0678	0662	0647	0632	0617	.0603	0589	0575
2 4	0561	0548	0535	0522	0510	0497	0485	.0473	0462	.0451
2 5	0439	0429	0418	0407	0397	0387	0378	0368	0358	0349
2 6	0341	0332	0323	0315	0307	0299	0291	0283	0276	0268
2.7	0261	0254	0247	0241	0234	0228	0222	0216	0210	0204
2.8	0198	0193	0188	0182	.0177	0172	0167	0163	0158	0154
2.9	0149	.0145	0141	0137	0133	.0129	0125	0122	.0118	.0115
3 0	.0111									

The area table will prove to be the more useful, however, and here are some of the uses to which it may be put.

1. It is consulted if we wish to find the number or proportion of cases in a normal distribution which lie on one side of a point along the scale.

Example: An I.T. set of scores have a mean of 100 and S.D. of 15. Find the percentage of scores which lie above 120.

This score of 120 is 20 above the mean

or in terms of sigma-scores $\frac{20}{15}$ or 1·333 above the mean.

From the table we see that $\frac{x}{\sigma}$ value of 1·33 gives a percentage of 40 82·for the area between the mean ordinate and the given one. (By interpolation we get the value of 40·88 for 1·333.)

As the curve is symmetrical about the mean ordinate 50% of its area lies above (to the right of) this line.

Thus the percentage of scores which lie above 120 is $(50 - 40{\cdot}88)\%$ $= 9{\cdot}12\%$.

To convert this to an actual number we should multiply the total number of cases by $\frac{9\cdot12}{100}$.

2. It is easy to extend the above to find the percentage of or number of cases which lie between two points on the scale. The process outlined in (1) is repeated in respect of both points and a simple subtraction gives the required result.

3. The table may also be used to find the point on the scale above or below which a given number or percentage of the cases in a normal distribution lie. This is the reverse of (1).

Suppose 15% of the cases lie above the required point. Then, considering only one side of the curve $(50 - 15)\%$ or 35% of the cases will lie between it and the central line. We therefore search in the body of the table to find an $\frac{x}{\sigma}$ value corresponding to this.

The value is therefore 1·036 (by interpolation) and if $\sigma = 15$ the required point is $1{\cdot}036 \times 15$ along the x axis.

If the mean is given by 100 this point will be $100 + 1{\cdot}036 \times 15 = 115{\cdot}5$.

This type of calculation may be extended to find the x distance on each side of the mean which cuts off a certain middle proportion of the cases. We can divide this proportion by a half and work on one side of the mean only, thus taking advantage of the symmetrical properties of the curve.

4. The curve may also be used for finding certain probable values and for obtaining an understanding of what is meant by probable error. There are various arithmetical ways of expressing a probability. If we say that 'it will probably rain tomorrow' we mean that the chances of rain are greater than those that it will

keep fine, that is, slightly more than the 1 : 1 or even chance. The probability is rather more than $\frac{1}{2}$ or 50%. In the case of the 'normal curve', probabilities are measured as ratios or percentages of a particular area compared with that of the whole. If the ratio or percentage is a small one the probability is correspondingly small. For example, a probability of $2\frac{1}{2}$% would be 1 chance in 40; a probability of 98% would be 49 chances in 50. Statistics is full of probabilities and the student should try to think in these terms. Probabilities are not certainties but refer to what is likely to happen in the long run and with a sufficiently large number of cases. Even though the chances that an event will happen or that a result is significant may be very much greater than the chances that the event will not happen or that the result is not significant, there is still an uncertainty. Many of the so-called 'laws of science' are to be thought of as being true to the extent of a large probability based on the results of a great number of observations. Probabilities of a sequence of chance happenings are subject to the rules of the behaviour of a single happening and no further prediction can be made. For instance, if we toss a penny four times and four successive 'heads' result, the probability that we shall throw a 'tail' on the fifth toss is no greater nor less than it was at the start. It is still an 'even chance', i.e. a probability of $\frac{1}{2}$ or 50%.

Suppose that the curve represents 'errors' or deviations from the mean. If we divide the area of the curve into halves by taking the 'middle' half of the scores we shall have 25% of the measures on each side of the mean line. The chances are even that any measure selected at random will lie within the 'middle' half of the scores.

We can find the distance of the x value which marks the boundary of the 25% of area by consulting the table. A rough value is $\frac{x}{\sigma}$ is .67, but by interpolation or by consulting a book of statistical tables we can obtain a more accurate value. We find that the chances are even (the probability is $\frac{1}{2}$) that any measure, score or error selected at random from a normal distribution will deviate from the mean by more (or less) than $.6745\sigma$.

TABLE III

PER CENT OF TOTAL AREA UNDER THE NORMAL CURVE BETWEEN MEAN ORDINATE AND ORDINATE AT ANY GIVEN P.E. DISTANCE FROM THE MEAN[1]

$\frac{x}{P.E.}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0	00	27	54	81	1 08	1 35	1 61	1 88	2 15	2 42
1	2 69	2 96	3 23	3 49	3 76	4 03	4 30	4 56	4 83	5 10
2	5 37	5 63	5 90	6 16	6 43	6 70	6 96	7 23	7 49	7 75
3	8 02	8 28	8 54	8 81	9 07	9 34	9 59	9 85	10 11	10 37
4	10 63	10 89	11 15	11 41	11 67	11 93	12 18	12 44	12 69	12 95
5	13 20	13 46	13 71	13 96	14 22	14 47	14 72	14 97	15 22	15 47
6	15 71	15 96	16 21	16 46	16 70	16 95	17 19	17 43	17 68	17 92
7	18 16	18 40	18 64	18 88	19 12	19 35	19 59	19 82	20 06	20 29
8	20 53	20 76	20 99	21 22	21 45	21 68	21 91	22 14	22 36	22 58
9	22 81	23 03	23 25	23 48	23 70	23 92	24 13	24 35	24 57	24 79
1.0	25 00	25 21	25 43	25 64	25 85	26 06	26 27	26 48	26 68	26 89
1 1	27 09	27 30	27 50	27 70	27 90	28 10	28 30	28 50	28 70	28 89
1 2	29 09	29 28	29 47	29 66	29 85	30 04	30 23	30 42	30 60	30 79
1 3	30 97	31 15	31 34	31 52	31 70	31 87	32 05	32 23	32 40	32 58
1 4	32 75	32 92	33 09	33 26	33 43	33 60	33 76	33 93	34 09	34 25
1 5	34 42	34 58	34 74	34 90	35 05	35 21	35 36	35 52	35 67	35 82
1 6	35 97	36 12	36 27	36 42	36 57	36 71	36 86	37 00	37 14	37 28
1 7	37 42	37 56	37 70	37 84	37 97	38 11	38 24	38 37	38 50	38 63
1 8	38 76	38 89	39 02	39 15	39 27	39 39	39 52	39 64	39 76	39 88
1 9	40 00	40 12	40 23	40 35	40 46	40 58	40 69	40 80	40 91	41 02
2 0	41 13	41 24	41 35	41 45	41 56	41 66	41 77	41 87	41 97	42 07
2 1	42 17	42 27	42 36	42 46	42 55	42 65	42 74	42 84	42 93	43 02
2 2	43 11	43 20	43 29	43 37	43 46	43 54	43 63	43 71	43 80	43 88
2 3	43 96	44 04	44 12	44 20	44 28	44 35	44 43	44 50	44 58	44 65
2 4	44 73	44 80	44 87	44 94	45 01	45 08	45 15	45 21	45 28	45 35
2 5	45 41	45 48	45 54	45 60	45 67	45 73	45 79	45 85	45 91	45 97
2 6	46 03	46 08	46 14	46 20	46 25	46 31	46 36	46 41	46 47	46 52
2 7	46 57	46 62	46 67	46 72	46 77	46 82	46 87	46 91	46 96	47 01
2 8	47 05	47 10	47 14	47 19	47 23	47 27	47 31	47 36	47 40	47 44
2 9	47 48	47 52	47 56	47 59	47 63	47 67	47 71	47 74	47 78	47 81
3 0	47 85	47 88	47 92	47 95	47 98	48 02	48 05	48 08	48 11	48 14
3 1	48 17	48 20	48 23	48 26	48 29	48 32	48 35	48 37	48 40	48 43
3 2	48 46	48 48	48 51	48 53	48 56	48 58	48 61	48 63	48 65	48 68
3 3	48 70	48 72	48 74	48 76	48 79	48 81	48 83	48 85	48 87	48 89
3.4	48 91	48 93	48 95	48 97	48 98	49 00	49 02	49 04	49 05	49 07
3 5	49 09	49 10	49 12	49 14	49.15	49 17	49 18	49 20	49 21	49 22
3 6	49 24	49 26	49 27	49 28	49 30	49 31	49 32	49 33	49 35	49 36
3 7	49 37	49 38	49 39	49 41	49 42	49 43	49 44	49 45	49 46	49 47
3 8	49 48	49 49	49 50	49 51	49 52	49 53	49 54	49 55	49 56	49 57
3 9	49 57	49 58	49 59	49 60	49 61	49 61	49 62	49 63	49 64	49 64
4.0	49 65	49 66	49.67	49 67	49 68	49 68	49 69	49 70	49 70	49 71
4 5	49 88									
5 0	49 968									
5.5	49 9896									
6 0	49 9974									
7.0	49 99968									
8 0	49.9999966									

[1] $\frac{x}{P.E.}$ is distance along x axis divided by probable error.

·6745σ is called probable deviation, and a probable error is ·6745 × standard error.

A third table gives the areas of the normal curve under certain values of x expressed in terms of probable deviation instead of standard deviation (sigma σ) values. As is to be expected, 25% of the area on either side of the central line gives an $\frac{x}{\text{P.E.}}$ value of 1.

Fitting a Normal Curve to a Series of Measures given in the form of a Frequency Polygon

It is better to draw the histogram or frequency polygon on graph paper to a suitable scale so that the paper is comfortably filled. The S.D. of the measures should be calculated after they have been grouped into frequencies.

(1) The height of the normal curve (see Appendix III) may be calculated from

$$y_0 = \frac{N}{\sigma\sqrt{2\pi}}$$

when N is the number of measures and σ is the standard deviation.

(2) The mid-point of each interval should be calculated in terms of sigma units by dividing each x value by the standard deviation.

(3) By using Table II the heights of the ordinates at each of these points is calculated. The table gives these values as a proportion of this ordinate and the actual heights are found by multiplying the height of the normal curve (mean ordinate) by the figure found in the table. The curve may then be plotted by joining the tops of the vertical ordinates with a smooth curve.

Inevitably there will be discrepancies between the actual ordinates and those obtained from the perfect curve. The sum of the theoretical frequencies of the curve should always be slightly less than those of the given distribution. The probability that a given distribution has discrepancies (which make it differ from a theoretical distribution) which are not due to chance can be found by using Chi-squared and consulting the appropriate tables.

The curve has some other uses in educational statistics. It can be used for setting standards for the distribution of marks, to assign values of difficulty to questions in a test, to give numbers of pupils in equal ability or talent ranges, for making scales for measuring various factors in addition to those of a purely cognitive type. It is often convenient to consider the curve as extending from -3σ to $+3\sigma$ or even from -2.5σ to $+2.5\sigma$ only. The student will bear in mind the nature of the small errors so introduced.

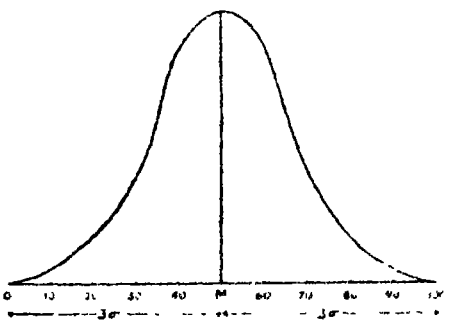


Fig 16

CHAPTER VI

MARKING AND ITS PROBLEMS

It is both amusing and disturbing to think that in many schools and colleges, lists of marks which have been produced in an arbitrary and entirely unscientific manner are thought to have an absolute value which bears no relation to the means by which they are obtained. For weal or woe no small part of the work of many teachers is the production of mark lists and the compounding of marks. It is well to give a little thought to the foundations of our beliefs concerning marks, particularly when these have been regarded as sacrosanct and as a type of numerical label by which one individual differs from another. A moment's thought will serve to show the limitations of certain marking systems. It would be a bold man who in marking two essays would give thirteen marks out of twenty to one and fourteen to another and be certain that the second was 5% better than the first! It would be a still bolder man who insisted that he was sure, in an English examination of the old type, that a candidate with 96 marks out of 100 was 1% better than another with 95 marks.

We can begin by summarizing the chief uses of marking systems:

1. *To obtain an order of merit list*

This is the popular use of marks in the schools. In order that there shall be a good spread it is necessary to devise a test which will give a normal distribution of the marks, or something approaching it. If two pupils have the same mark they will occupy the same place and the next pupil in order of merit will have the next but one place. If the mark list in order of merit is to be used for correlation purposes either by Spearman's method of ranks or by the 'footrule' it is wise to consider more carefully these 'tied' places.

e.g. The following is a portion of an old school mark list.

	Mark	*Position in rank*
Thompson	92	1
Allen	84	2
Walker	81	3 =
Smith	81	3 =
Brown	81	3 =
Jones	79	6
Turner	76	7

In this case it is better to credit Walker, Smith and Brown with the average place, i.e. the fourth place. In the same way, suppose two boys 'tie' in the mark which comes after the 10th place. Instead of putting two 11th places between the 10th and the 13th places it is wise to credit the two boys with equal marks with '11½' places each. If correlation is to be performed this is particularly important.

2. *To separate candidates who reach a certain level from those who do not*

Most of the public examinations, such as those for School Certificate, matriculation, degrees and diplomas have this end in view. At first sight this may seem easy, but it is beset with pitfalls. It is unwise to draw our lines of demarcation on the frequency curves at points where the curve is at its highest, for here there is less chance of a critical separation of one class of candidates from another. The standard of examination papers and of students taking the examination varies from year to year. It is difficult or impossible for an examiner who has set an examination paper to know what standard it is by just looking at it. Only experiment with many trials will show, and this is not usually possible. Examiners are changed from year to year or after a short period of years. Many examining bodies 'standardize' the marks, by approximating the percentages of credits, passes, failures and even distinctions respectively from year to year. It follows that in a year when many good candidates present themselves it is much more difficult to pass the examination than when there are more weaker candidates.

3. *Tests and examinations may be set by a teacher to test the value of his own work or to estimate the progress already made by a class*

This should help the teacher to find what is difficult and what is easy to the pupils in his own teaching, and he can amend his work accordingly.

4. *Examinations should also look forward*

and not only backward on the pupil's past work. In other words, examinations should be prognostic. How far they have this quality has been the subject of considerable investigation. If the boy or girl at eleven has reached a certain standard in Arithmetic and English is he or she a fit candidate for a place in a grammar school? Entry to the old universities may be secured with scholarships if a candidate shows sufficient knowledge of and ability in Mathematics. Is this a sufficient guarantee of a satisfactory university and subsequent career?[1]

Examinations are not as reliable as they ought to be for some or all of the following reasons:

(1) The number of questions of the older or essay type which the candidate is able to answer in the allotted time is so small that there is insufficient sampling of the candidate's knowledge. Questions of 'luck' or 'chance' figure too largely in the result, from the candidate's point of view.

(2) Candidates may differ in mental and physical condition from day to day and this will affect performance in the examination. Vitamin intake, digestion, hours of sleep, mild infection, other physical and emotional states, the time of day, atmospheric and other environmental conditions and the total length of the examination may modify the student's work in it, or in some part of it.

(3) Particularly in the 'Arts' subjects there may arise differences of opinion between one examiner and another concerning the value of a student's work.

(4) Examiners are not always consistent with one another in

[1] An excellent short examination of examinations is given in Chapter XI of P. E. Vernon's *The Measurement of Abilities*.

their standards of marking. Nor will the same examiner adhere to the same standard at different times of the same day, at different parts of the week and at different stages in marking a large batch of examination papers.

The compounding of marks is a still more difficult task. Here the idiosyncrasies of a number of markers in different subjects will produce anomalies in the final result which are both unfair and misleading. As so much is often made to depend on the sum total of a candidate's achievement in an 'omnibus' examination, it is the duty of all concerned in the matter to investigate carefully what really lies behind the masses of figures which are produced from the several subjects of the examination.

In a public examination, such as the Intermediate Examinations of the University of London, it may be possible to give equal weight to each of the subjects which are taken; but in a school annual examination this is not possible, nor is it for the marks which are given on each term's work. It is obvious that the maximum marks in English should be greater than those for Geography, just as those in Mathematics will usually be greater than those in Chemistry. The reason is the obvious one that more hours per week are devoted to English than to Geography, to Mathematics than to Chemistry. (We will leave the problem of relative importance from other points of view, though few would contest the superior position of English in the school curriculum.)

A reasonable way of treating the marks of the respective subjects before compounding them would be to arrange each maximum mark so that it is proportional to the time devoted to the particular subject each week.

Suppose 5 hours are spent on English, 4 hours on Mathematics, 3 hours on Science and 2 hours on History. We might allow a term's maximum of 200 marks for English, 160 for Mathematics, 120 for Science and 80 for History. It may happen that the total for all subjects will come to some large number which is not a multiple of a hundred. Whatever the total maximum, that is, the total of the maxima of all the subjects, an order of merit can be found just as easily, and if a percentage is required of the maximum score this can subsequently be found by simple reduction.

There is usually a more serious difficulty in the compounding of marks. Some markers feel that a normal distribution of marks tends to depress and discourage all but the top quartile division of the candidates, whilst others feel that they may force their students to strive for better ultimate examination results by marking stiffly the work and tests of the term. Again, others find marking so difficult that they are only able to separate from the mass of papers the very poor candidates and the very good ones, and all others are bunched together with very little spread or dispersion of marking and a rather high average usually of about 55%. This makes the compounding of marks difficult. We can do something to adjust the various marking scales which will improve matters somewhat. Each mark may be regarded as a positive or negative deviation from the mean which is called o, or the marks may be standardized by dividing these deviations by the standard deviation. All this would involve much labour which would certainly not be welcome and might not be possible at the end of term. The marks might be improved for the purposes of compounding by adjusting the marks in the interquartile range by means of a graph.

Another useful expedient is to adjust the marks by means of a straight-line graph so that the top boy gets the maximum marks and the bottom boy no marks. (The objection to this is that the top boy may not be worthy of the maximum marks just as the bottom boy will probably deserve something better than zero marks.) All the objections in theory are met, however, by the very practical result that the resulting order of merit is much fairer to all concerned. We have said enough to show that no system of marks is entirely above criticism, and if we keep in mind the difficulties of marking and compounding our marks our system will progressively improve.

Most teachers soon evolve a personal system of marking, and it is well for all who have to mark the work of pupils and students to explore the fundamentals of their own ideas on the subject. It is more difficult to mark papers of the essay type than those of the new style where there are many shorter questions, which usually only require a sentence in answer to each, or even the

choosing of a correct word or sentence from a number which are given for each question. It is obviously more difficult to mark an English essay where style is taken into consideration than an arithmetic paper where a marking scheme can be followed fairly closely. Where marks are deducted for errors, markers should see that the total reduction bears a reasonable relationship to the marks credited for correct work. Until much practice has been obtained in marking and the marker has subjected his work to careful examination, it will be inevitable that a careful re-marking of a batch of papers, after a first assessment, will be desirable. This will enable the earlier papers in a batch to be adjusted to those which have come later and have been marked in 'a state of maturity' for that particular examination. Some conscientious examiners arrange the papers in order of merit as shown by their marking and then re-read them in descending order of merit, satisfying themselves that each paper is a little less worthy than the one which preceded it. If the examination and the candidates have been fairly matched the marks should be distributed in a normal manner or in an approximation to it. In the case of fairly homogeneous small groups (e.g. the mathematical 'sets' of a large fifth form) it is difficult to obtain the requisite distribution of the marking. It is obvious that the larger and more heterogeneous is the group the easier will it be to obtain normal distribution. It may be allowable in a scholarship examination when only a very few of the finest candidates can obtain awards to permit a slight positive skew to the distribution and thus give a better spread in the upper reaches of the marking. In the same way it may be permissible to allow a little negative skewing if the intention of the examination is merely to reject a few candidates who fail to secure a minimum of marks less than 40% or 50%, but the fact remains that for general purposes normal distribution should be aimed at and the marks which separate one class or degree of merit from another should not coincide with the mode (which in the case of normal distribution would also equal the mean and the median).

A simple problem in connection with marking is the reduction of marks. The marks have been given to one maximum mark and it is desired to reduce or translate them to another scale with

a different maximum. It is presumed that it is not desired to interfere with, or endeavour to modify in any way, the relative distribution of the marks which would be best achieved by drawing a curved-line graph.

The simple task of 'reducing marks' is best effected by one of three ways:

(1) Using a slide-rule
(2) Drawing a straight-line graph.
(3) Multiplication of the marks by an easy fraction.

1. Using a slide-rule.[1] This simple instrument permits multiplication and division sums to be performed by adding or subtracting lengths of a ruler. As the standard engineer's slide-rule permits the use of various functions and is a more complicated instrument than we require for the simple reduction of marks, some schools possess a large slide-rule which is graduated for multiplication and division only. Suppose we have marked to a maximum of 120 marks and we wish to reduce these marks to a maximum of 100, that is, to express them as a percentage of the maximum. We take the slide-rule and move the lower scale (B) so that the graduation 12 on it corresponds with 10 on the upper scale (A). The given mark is found on scale B and the reduced mark is read opposite to this on scale A.

2. A 'ready-reckoner' table can be made in convenient form by drawing a straight-line graph. It is best to use graph paper where each large division contains ten (and not five) small divisions for this will facilitate reading the graph. To take the case given above. A point on the graph paper, on which axes have been drawn horizontally at the bottom of the paper and vertically on the left side, is found which corresponds to the maxima in the given and on the reduced scale. This will be the point with x value 120 and y value 100. The point 12·10 (counting in large squares) is found and joined to the point 0 (the point of intersection of the axes) and the resulting straight line is the graph required. It is only necessary to find the corresponding y value on it when an x value (that is, a mark on the 120 maximum scale) is read off horizontally.

[1] See Appendix II.

3. Many simple reductions can be performed by rapid mental arithmetic. Reductions to a half or a tenth or by two-thirds, a fifth and so on would give no trouble. A reduction which frequently occurs is from 25 to 10 as maxima. This is equivalent to dividing by $\frac{5}{2}$ which is equal to $\frac{10}{4}$. Thus we divide each mark on the 25 scale by 4 and multiply it by 10 by shifting the decimal point one place to the right. The reduction from a maximum of 120 to one of 100 is equivalent to multiplying by the fraction $\frac{5}{6}$ or $\frac{125}{150}$.

Most people could achieve this very quickly by adding a nought to each mark on the 120 scale to multiply it by 10 and then dividing each number by 12. Some conscientious teachers who find difficulty in handling figures obtain their reductions by one method and check them with another.

The importance of the transfer examination which is now taken by all children in state-controlled schools at the end of their primary school life has become greater, not less, since the passing of the Education Act of 1944. In view of the fact that the whole subsequent life and career of a child may be modified by the type of secondary education which he receives, it is hardly necessary to say that anything which can be done to improve the transfer examination, which is taken at about the age of eleven, should be regarded as a matter of prime importance. We should look upon the test as one which should have a prognostic value. Although statistical analysis in these matters is probably of less importance than the sound framing of the test papers, it is only by mathematical investigation that we can be assured that we are on the right lines in our examination methods. Much yet remains to be done, but all honour should be given to Professor Godfrey Thomson, who has devoted many years of his life to these problems and with his staff has evolved the Moray House Tests. It is obvious that the standard of the tests should be maintained from year to year and that the tests should aim at determining the type of secondary education which will best fit a particular child rather than testing the attainment and factual content of the child. Accordingly, tests in English, Arithmetic, and a General Paper which seeks to explore the native capacity (often called intelli-

gence) of the child, are prepared with this end in view and are standardized by exhaustive experimental tests. It is easy to imagine that ideal tests for children of 10-11 cannot be evolved by 'an armchair process', but only painstaking trial and error and careful analysis of the results will suffice. Even so, no ideal tests have yet been found, and there is still at least 10% error in the *prognostic* value of most transfer tests. Nor is the underlying psychological theory a matter on which there is complete agreement between eminent authorities. It is believed that the average verbal ability of girls at the transfer age is somewhat greater than that of boys. This difference would appear to be accentuated in the case of children from the country, that is, from small villages rather than from towns. Dr W. P. Alexander has stressed repeatedly and with justification the necessity of allowing for non-verbal abilities in transfer examinations, and he would divide abilities by means of oblique factors[1] into verbal and non-verbal types. Enough has been said to show that the serious student interested in the transfer examination will find much data which can be explored by statistical methods and will yield useful results. These must still be regarded as being valuable even when they only serve to show us the weaknesses of our methods and do not always offer any ideas for their improvement.

In connection with transfer examinations and attainment tests an important matter susceptible to statistical treatment is the age allowance in marking schemes.

Some education authorities permit only a single attempt at a transfer examination, and there is thus an age range of a year. Allowance is made for differences of not less than a month. Other authorities have an age range of two years or even more and permit two attempts at the examination if necessary. In fixing this allowance it is wise to make experiments with large numbers of children of various age groups and to use general papers containing tests of 'intelligence', English and Arithmetic rather than papers of more limited scope. We could set a series of papers to children in age groups of 12, 11 and 10 respectively, and find the median score for each paper (or set of papers) and for each

[1] See page 109.

group. The median score or norm would show an increase when using the same paper from year to year. By drawing a graph for each paper (or set of papers), using the three points of the 12, 11 and 10 norms, we find that we can find a straight line which practically goes through the three points in each case. If we use the graph to call the 12-year norm 100, we can read off the 11-year and 10-year norms on this scale. The graphs obtained from the median scores of the other sets of papers will have different slopes, but when the 12-year median score is called 100 and the other norms multiplied by the same fraction or read off on the graph we shall probably find that the other norms differ a little for the same age group. The average is then taken.

Suppose that the difference averages about 2½% per year. At first sight it may appear that 2% should be added to the marks of the candidates for every month of his age below 12 years. This would probably be unfair as 2% of a lower mark is obviously less than that of a higher. To overcome this several methods are employed. We can take the age of the pupil with the greatest number of marks and reckoning two marks per month as an age-allowance scale up his marks to those which would be expected if he were 12 by means of a graph or a slide-rule. The corrections work out as follows:

Age	*Per cent*
12·0	100
11·11	98
11·10	96
11·9	94
11·8	92
11·7	90
11·6	88
11·5	86
11·4	84
11·3	82
11·2	80
11·1	78
11·0	76
etc.	etc.

Thus we should find the percentage corresponding to the age of the pupil and multiply his marks by a fraction with this percentage in the denominator and 100 in the numerator. e.g. Suppose a boy of 11 years 4 months obtains a total of 362 marks. His expected achievement at the age of 12 years precisely would give

$$362 \times \frac{100}{84}$$

or 431 marks.

The matter may be regarded from another angle: we have obtained norms for each age group and by interpolation we can obtain norms for each month. Every pupil's marks will correspond with a particular age norm and therefore we could give an assessment of the achievement of each pupil in terms of his test or examination age, that is, the number of months above or below average as an equivalent of a greater or lesser ability than the normal for his age.

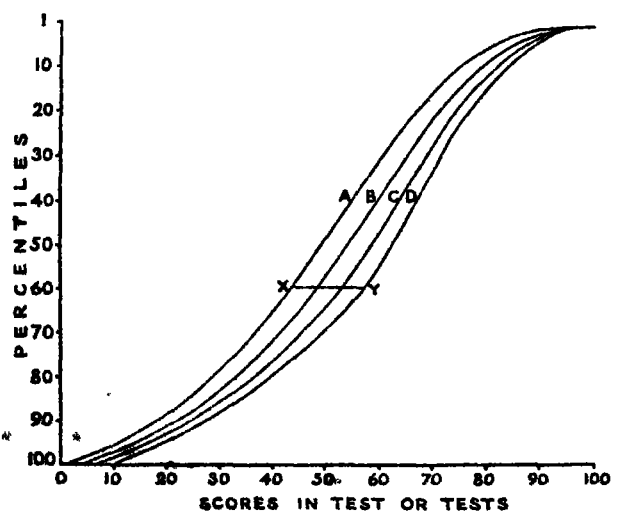


Fig. 17. Percentile curves for four three-month groups. XY represents an age allowance for 9 months at a particular percentile level. This level must then be interpreted in terms of the scores of the whole of the candidates from a separate curve. For convenience percentiles have been reckoned from the highest score.

In transfer examinations some authorities, following a method similar to that which we have outlined above, have a table of percentages of marks which are added to the total scores of the children according to their ages. A cruder method which is employed by others is to have a table of marks and add an appropriate number to those of a child in regard to his age but without regard to his achievement. Strictly speaking, the percentage or proportional way of making the increase is the only equitable way, for the method of adding fixed numbers of marks according to age benefits the weaker children at the expense of the more able.

The best method for ordinary use and one which does not evolve a great deal of labour is that due to Thomson.[1] The total marks (or those in separate subjects) for every child are divided into four age groups 11.0 years to 11.2 years inclusive, 11.3 to 11.5 years, 11.6 to 11.8 years, 11.9 to 11.11 years. Cumulative frequency (percentile) curves are drawn for the marks in each group. The abscissae differences between the first and the fourth curves give the differences in marks corresponding to a 9 months' age difference. It will be noted that this difference is one of 9 months and not of 12 months as each curve is for the average age of the three-month age group, that is, the first curve is centred on an age of 11 years 1½ months and the last on 11 years 10½ months. It is now necessary to interpret these in terms of the percentiles and marks of the whole 11-year group taken together. Usually no child under 11 is given more than the allowance for 11.0 years. The mark difference for 9 months is divided by 9 to give the monthly adjustment for each score level. Equivalent marks are subtracted for children from 12.0 years to 12.11 years.

There remains the question of the ideal mark scale and the mark value of each question in a given test. These matters can best be understood by further reference to our curve of normal distribution. It will be seen that if we draw vertical lines at distances of 3σ on each side of the central point the area enclosed by these lines and the curve is practically the whole of its area. Now, the area of the curve gives the frequency or the number of

[1] See *The British Journal of Educational Psychology*, 1936.

TABLE OF AGE ALLOWANCES IN A TRANSFER EXAMINATION

Percentile	*Total Score*	*Difference for 9 months*	*Monthly Difference*	*Allowance of Marks* (ADD for 11+ group)												
				11.11 11	10 11	9 11.8 11	7 11.6	11.5	11.4	11.3	11.2	11.1	11.0			
72	106-122	15 5	1.7	0	2	3	5	7	9	10	12	14	16	17	19	21
63	123-140	16	1.8	0	2	4	5	7	9	11	13	14	16	18	20	22
52	141-157	17	1 9	0	2	4	6	8	10	11	13	15	17	19	21	23
44	158-175	16	1 8	0	2	4	5	7	9	11	13	14	16	18	20	22
35	176-192	15.5	1.7	0	2	3	5	7	9	10	12	14	16	17	19	21
27	193-210	14	1 6	0	2	3	5	6	8	10	11	13	14	16	18	19
19	211-227	13	1 4	0	1	3	4	6	7	8	10	11	13	14	15	16
12.4	228-245	12	1.3	0	1	2	4	5	6	8	9	10	12	13	14	15
7.4	246-262	10.8	1 2	0	1	2	4	5	6	7	8	10	11	12	13	14
3.5	263-280	8	.9	0	1	2	3	4	4	5	6	7	8	9	10	11
1.1	281-297	6.3	.7	0	1	1	2	3	3	4	5	6	6	7	8	8
.3	298-315	2	.2	0	0	0	1	1	1	1	1	2	2	2	2	2

12.0	12.1	12 2	12.3	12.4	12.5	12 6	12.7	12.8	12.9	12.10	12.11

(SUBTRACT for 12+ group)

Here the percentile orders have been reckoned from the highest score as it was not necessary to deal with candidates beyond a percentile level of 72 counting from the highest score This is for convenience and care should be taken to see that it does not lead to confusion in describing percentile levels.

cases or scores, and only ·2% of the scores lie beyond the 3σ lines at the left and right extremes of the curve. (This will be clear from our short chapter on the normal curve.) If instead of drawing our vertical lines at points 3σ from the centre we choose points at a distance $\frac{5}{2}\sigma$ on each side of this point, the area of the curve thus enclosed is 98·76% of the whole, that is to say, we have omitted only 1·24% of the whole scores. Although we have made slight sacrifices to accuracy it is very convenient to have a base of 5σ instead of 6σ because we can more readily divide it into a ten- or a hundred-part scale, and for our purpose here this arrangement is quite accurate enough

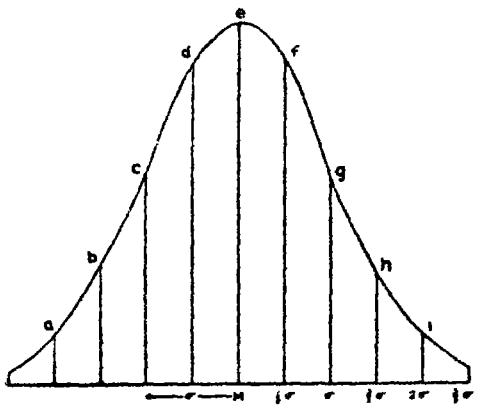


Fig 18

Suppose now that we divide it into 10 equal divisions along its base, and further let us imagine that in a test we have this number of properly graded questions, so that on drawing a graph showing the number of persons solving each question we get a distribution curve of the normal type.

The scale of ability is taken to be similar to that of the scale o difficulty of the questions. Now area '*a*' is equivalent to th number of those who cannot solve Question 1. Similarly are '*ab*' represents the number of those who cannot solve Question 2 '*abc*' those who cannot solve Question 3, and so on. Obviousl

the mark value of a question should increase with the proportion of people who fail to solve it. For instance, by consulting the tables giving the proportions of curves of normal distribution which are cut off by ordinates at particular distances from the central point,[1] we can find that the area *abcdefg* is approximately 85% of the area of the whole curve. Hence Question 7 would be too hard for 85% of the candidates but it could be solved by the remaining 15% (assuming that the time factor did not enter).

Thus if a question is solved by 15% of the candidates it will be of difficulty 7 and take this number of marks.

We can take the matter a step forward by drawing a percentile curve showing the percentages of candidates failing to solve each problem according to its difficulty and the marks which will be given to it.

The student will find the construction of such a curve and the following tables an easy exercise in the use of the normal distribution or probability-integral tables:

Marks per question	*% able to solve it*	*% failing to solve it*
1	98·35	1·65
2	94	6
3	85	15
4	70	30
5	50	50
6	30	70
7	15	85
8	6	94
9	2	98
10	almost 0	almost 100

In order not to break too much with time-honoured custom and yet maintain a system which permits a mathematically reliable compounding of marks, some authorities regard 90% as the highest mark and 30% as the lowest in all but exceptional cases. Only one candidate in several hundred or even a thousand is regarded as being so excellent that he achieves more than 90%

[1] See page 70.

or so feeble that he scores less than 30%. This method, being used by schoolmasters and in certain of the public university examinations, obviously implies a certain degree of homogeneity resulting from the selection of the more able individuals from the population at large.

A reasonable dispersion would be given by a standard deviation of 10 and, assuming a normal distribution, a median of 60. In this case the percentages of candidates expected to achieve scores in various mark groups would be as follows: (The extreme upper and lower reaches of the marking are reserved for candidates of rare brilliance or poverty of achievement.)

Mark %	% *in each group*
92-88	up to ½ %
87-83	1
82-78	3
77-73	6½
72-68	12
67-63	17
62-58	20
57-53	17
52-48	12
47-43	6½
42-38	3
37-33	1
32-28	up to ½ %

In practice, things do not work out quite as easily as this. Marks have to be allowed in many cases for answers which are partly correct and in many tests a choice of questions has to be permitted. In the 'new-type' examinations the number of questions would be much larger than in the old type and answers would be *right* or *wrong*, for the most part. Also, in view of the larger number of questions, proper sampling of the candidates can be achieved and there is no need to permit selection on the part of candidates. Nevertheless, in any type of examination a proper order of merit will only be secured by a proper grading of questions in difficulty,

with a weighting of marks in accordance with the requirements of the curve of normal distribution. It is not pretended that practical achievement in examining can match up to theoretical ideal demands but a more careful mathematical analysis of each test will go far to improve a system of examinations which has not yet been replaced as a means of assessing ability and achievement.

In a work well known to the point of notoriety Hartog and Rhodes produced evidence to show the unreliability of examination. No doubt 'An examination of examinations' was intended to make our flesh creep, and to sustain their thesis the authors chose cases which did all they could to show the subjectivism of marking in the worst possible light. Most of the sets of scripts which were used for their experiments were more homogeneous than we should ordinarily find. Such sets of papers always present difficulties and it is well known that to secure a distribution which approaches a normal one we must use a large and heterogeneous group. Nevertheless, the work of these authors did much to bring a realization of the need for more care in examinations no matter at what level.

On the other hand, the value of examinations and the care and thought with which they are conducted has been finely expressed by Bidereton in *The Case for Examinations*. It is a step forward if only average marks and standard deviations or interquartile ranges are equalized between one examiner and another or between one subject and another before marks are compounded. There is an increasing awareness of the necessity of this, and that a failure to do so will lead to erroneous and anomalous results in final order of merit lists.

It must not be assumed that the new type of test is in all ways superior to the old, or that it is free from defect. Vernon in *The Measurement of Abilities* has given an excellent analysis of this matter. Much more time, skill and experience are necessary for the production of the new type test-paper containing many graded questions, but time is saved in marking the scripts. Unless the number of scripts exceeds 300 no time is saved on the aggregate of setting the papers and marking the scripts. The examiner must decide just which type of question suits his purpose for the subject matter in hand. The questions may be divided into the

following types: (*a*) Simple recall and 'open-completion', where blank spaces in the question have to be filled in. (*b*) True-false where there is a set of statements some of which are true and some false. The candidate has to indicate 'which is which'. (*c*) The Multiple-choice type, including best reason and matching items. In each case a number of alternative answers are given. One is correct and this is to be underlined by the candidate. (*d*) Re-arrangement type. Here a list of items which should fall into a unique order is given in the wrong order. The candidate must rearrange them to give the correct order.

In the new-type tests a certain number of correct answers in the recognition-type of test may be obtained by chance guessing This only means that the zero level in scoring is equivalent to a score which could be calculated as being the percentage of marks which might have been obtained by pure chance. The marks obtained may be corrected for guessing by using the formula. True score $= R - \frac{W}{n-1}$ where R is the total number right and W the total wrong and n the number of alternative answers provided for each question. It has been shown that the above correction only makes appropriate compensations for the average candidate. On the whole the effect of guessing is much less than the layman would imagine.[1]

Mental Ages and Intelligence Quotients

The *Mental Age* (M.A.) of a child as given by an intelligence test. Its *Educational Age* (E.A.) as given by educational tests is equal to the actual or *Chronological Age* (C.A.) of an *average* child with the same test scores. *Intelligence Quotient* is given by

[1] The system of marking at most musical festivals and competitions seems to be extraordinary. Even very poor efforts are not infrequently given upwards of 75% and the majority of candidates obtain more than 85%. This is obviously intended to hearten all candidates and to maintain enthusiasm for subsequent occasions. Nevertheless, the adjudicator's task is rendered difficult by this system and his final marks are perforce given by reference to an order of merit resulting from a quick consideration of the qualities which make one competitor or group slightly better than another. The adjudicator needs good experience, judgment and memory.

$\frac{\text{Mental Age}}{\text{Chronological Age}}$ and is often expressed as a percentage. At first sight these may seem to be a much simpler and more straightforward method of describing attainments or abilities than the use of percentile levels. There are some difficulties, however. To start with, the growth of intelligence and educational abilities are not regular year by year. The upper limits of achievement vary from child to child. After the age of eleven the intelligence-test scale becomes so unreliable and artificial that it is wise to abandon M.A. units from the age of 12 upwards. The proportional advancement or backwardness of a child whether in educational achievement or intelligence tend to increase with increasing age. The fractions $\frac{\text{M.A.}}{\text{C.A.}}$ (i.e. I.Q.) and $\frac{\text{E.A}}{\text{C.A.}}$ (i.e. E.Q.) keep reasonably constant for a number of years.

There is nothing absolute about a scale of intelligence 'norms', or the marking scale of an intelligence test. Unless all intelligence tests (in addition to all the other desiderata) are standardized as regards mean or average and standard deviation, statements of I.Q. measurements will be ambiguous. We can only say 'the I.Q. of Smith as measured by this or that particular test is x'. The Moray House Tests yield an average score of 100 and an S.D. of 15. The Stanford Binet tests were formerly believed to yield an S.D. of 15 but this is now known to be 16½. In fact the S.D.s of intelligence-test scores vary from 12 to 25 (with a mean score of 100). The matter can only be made accurate by expressing differences in achievement in standard deviation units (see page 26).[1]

We have left until last a short statement of the chief difficulty, and one which is perhaps not apparent at first. It is that of establishing age norms. It is practically impossible to take a sufficiently large sample which will represent all possible children of any age group. In primary school life it is perhaps possible if we cast a wide net to find groups which give us a fair sample of the total population, but even here it is difficult to allow for the children (either bright or dull) who attend private schools or those who

[1] This section should be followed up with Chapter X of Vernon's *The Measurement of Abilities*.

go to special schools. After the age of 11, with the children in various types of secondary schools the problem becomes even more difficult. There is still room in the field of simple research by teachers for experiments using intelligence tests with children of various ages, physical types, 'social' positions, localities. Although many hundreds of thousands of such tests have been given there is still no shortage of opportunities for their use. In rare cases it has been possible to test all the children of a certain age or from a certain locality but more often the best that can be done is to select them from as many schools as possible in different districts to give as wide a range of social and economic differences as possible.

To Standardize an Intelligence Test

If we could give the intelligence test to very large numbers of children in year groups of 10, 11 and 12 (making sure that each group is truly representative of *all* children of that age), we could plot the three averages as equally-spaced ordinates on a graph and join the points. This would yield a straight line sloping upwards and by interpolation we could read off the monthly norms.

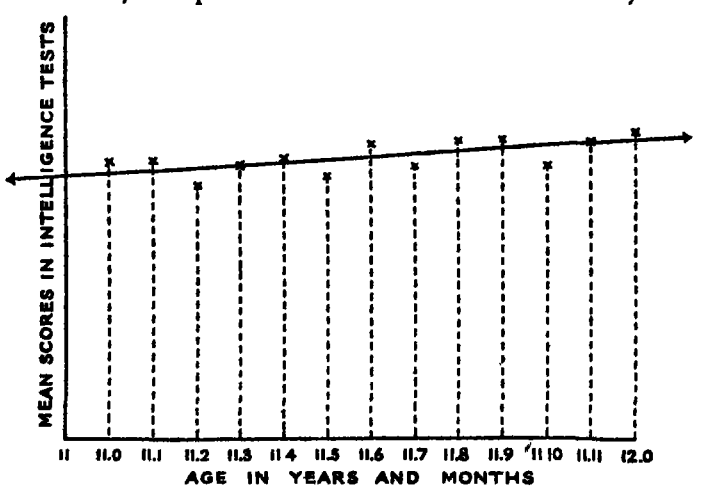


Fig. 19. The line of best fit is found by the method of least squares.

It would be convenient to have each of the ordinates separated by 12 units of abscissae in order to facilitate these monthly interpolations. This method would be open to many objections. The division into years is far too coarse and little attention is paid to finer differences in the 11+ year which may be the most important from our point of view, particularly if we are interested in the transfer examinations at the end of primary school life. Moreover, errors of sampling and distribution cannot be corrected by this method of taking the three year groups.

A much better method is that due to Thomson.[1] A 'complete, numerous and uncreamed' year is tested. The year group is divided up into 12 monthly groups, which must be as large and heterogeneous as possible so that each shall be a good sample of that age group of the whole population. The average score in the test for each monthly age group is found and plotted as an ordinate on a graph with abscissae giving the monthly spacings. Owing to errors in sampling the twelve (or thirteen) plotted points will usually not lie on a straight line. The line of best fit has to be found. As usual this is done by the method of least squares, that is, the sum of the squares of the deviations of the ordinate points from the line must be made a minimum.[2] The straight line of best fit can be extended backwards to deal with the 10+ age group and forwards for the 12+ group. A child's M.A. can therefore be read off on this line by reference to his score in the test. His I.Q. can be found by dividing by his chronological age.

Intelligence tests may also be standardized by comparing scores achieved in them with those in established tests such as the Binet, using the same groups of children.

[1] See *The British Journal of Educational Psychology*, 1932, page 99.

[2] The quantity $\Sigma(v^2)$ where the v's are the deviations from zero obtained when the twelve or thirteen points obtained from the scores are substituted in the equation of the straight line $y = mx + c$. The values of m and c which give this are found from the equations:

$$\Sigma(y) - m\Sigma(x) - nc = 0$$
$$\Sigma(xy) - m\Sigma(x^2) - c\Sigma(x) = 0$$

where x represents ages and y the scores.

CHAPTER VII

THE 'FACTORS' OF THE MIND

By measuring we know what things are long and what short. The relations of all things may be thus determined and it is of the greatest importance to measure the motions of the mind.

MENCIUS, *c.* 335 B.C.

In the early years of this century Professor Charles Spearman commenced a serious investigation into the nature of human abilities. 'One of the most pernicious (of fallacies) was found to be the current usage of the word "intelligence" without any definite idea behind it. Another, that does even greater mischief in practice, was the irrepressible tendency to assume that terms like 'attention', 'combination', 'analysis', 'range of association', 'co-ordination of hand and eye' and so forth represent so many functional unities or behaviour units. Alongside of these two great impediments to the advance of science has been the pseudo-explanation of the tests of a person's "intelligence" as measuring a 'level', 'average' or 'sample' of his abilities whereas really no measurement is conceivably possible.'[1] The works on educational psychology have persisted in telling us that the 'faculty' psychology is dead (which should be true) but there has been a tendency to resurrect it in terms of mental factors.

Spearman investigated five 'laws' quantitatively: the laws of Span, Retentivity (inertia and dispositions), Fatigue, Conation and Primordial Potencies (including such influences as those of age, sex, heredity and health). It was in these investigations in which he attempted to put certain aspects of psychology on a scientific basis that he made great use of correlation coefficients between tests, and examined them by mathematical analysis. At first it was necessary to achieve a 'Copernican revolution' in point

[1] C. Spearman, *The Abilities of Man*, pages 409-10.

of view. Instead of postulating 'an ill-defined mental entity the intelligence', and then by 'intelligence tests' trying to obtain a value for this, he started with a perfectly defined quantitative value '*g*' and then demonstrated what mental entity or entities this really characterizes.

Spearman showed that the coefficients of correlation between tests tend to fall into 'hierarchical' order and he further demonstrated that this was consistent with his 'Two Factor' theory.

An example will suffice to show how this works out:

Suppose the correlation coefficients between a number of tests 1. 2. 3. 4. 5. 6. are written down in rows and columns as follows:[1]

	1	2	3	4	5	6
1		r_{12}	r_{13}	r_{14}	r_{15}	r_{16}
2	r_{12}		r_{23}	r_{24}	r_{25}	r_{26}
3	r_{13}	r_{23}		r_{35}	r_{35}	r_{36}
4	r_{14}	r_{24}	r_{34}		r_{45}	r_{46}
5	r_{15}	r_{25}	r_{35}	r_{45}		r_{56}
6	r_{16}	r_{26}	r_{36}	r_{46}	r_{56}	

The tests which give each correlation ratio are denoted by the subscripts of *r*. e.g. r_{34} is the correlation coefficient between tests 3 and 4. The above arrangement of rows and columns is known as a MATRIX and in research work on psychological tests the elementary properties of such sets of numbers are of prime importance.

[1] The coefficient of correlation between two sets of measures is the proportion of the total variance which is due to the common factor in each test

$$r = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_a^2} = \frac{\sigma_c^2}{\sigma^2_{c+a}}$$

where σ_c^2 is the variance due to the common factor and σ^2_{c+a} the total variance. Note that variance is the square of the standard deviation and that variances may be added algebraically.

The exact nature of the tests in *this* case is of secondary importance Examples would be: Analogies; Opposites; Resemblances; Understanding instructions; 'Completion'.

Let us consider the matrix rewritten with numerical correlation coefficients:

Test	1	2	3	4	5	6
1		·48	·24	·54	·42	·30
2	·48		·32	·72	·56	·40
3	·24	·32		·36	·28	·20
4	·54	·72	·36		·63	·45
5	·42	·56	·28	·63		·35
6	·30	·40	·20	·45	·35	
Total	1·98	2·48	1·40	2·70	2·24	1·70

We have added up the coefficients in columns and now proceed to rearrange the matrix so that the totals of the columns are in descending order of magnitude thus:

Test	4	2	5	1	6	3
4		·72	·63	·54	·45	·36
2	·72		·56	·48	·40	·32
5	·63	·56		·42	·35	·28
1	·54	·48	·42		·30	·24
6	·45	·40	·35	·30		·20
3	·36	·32	·28	·24	·20	
Total	2·70	2·48	2·24	1·98	1·70	1·40

In this ideal case[1] the 'hierarchical order', as Professor Spearman called it, is easily seen. The correlation coefficients in any two columns have a constant ratio to one another. Consider the last two columns:

·45	·36
·40	·32
·35	·28
·30	·24
	·20
·20	

[1] Given by G. H. Thomson, *The Factorial Analysis of Human Ability*. (The hypothetical coefficients have been chosen to demonstrate the principle in the easiest way).

Ignoring those coefficients which are not paired it is easily seen that there is a ratio of 5 : 4 between the left and right columns. In other words each coefficient on the right is $\frac{4}{5}$ of that on the left.

This precise relationship would not be apparent in actual tests but the tendency would still be evident. Spearman explained this hierarchical order by a common factor '*g*' which was present in each test but in the largest quantity in that at the head of the hierarchy. Each test also contains a specific factor which would not be found in any other test unless similar varieties of the same test had been used. A test is said to be 'saturated' or 'loaded' with *g* to an extent depending on its place in the hierarchy. Suppose it were possible to devise a test of pure '*g*', that is to say, one completely saturated with '*g*' and containing no specific or '*s*' factor. Such a test would stand at the head of hierarchy. The self-correlations of the tests are ideally unity and in the diagonals of the matrices have been left blank. In the case of the self-correlation of pure '*g*' it can be written in and this number (unity) will conform to the hierarchy. In the other unities the 'specifics' enter and they are omitted as they do not conform to the rule of proportionality between the columns.

We may now rewrite the matrix including 'pure' *g*:

	g	*a*	*b*	*c*	*d*	*e*	*f*
g	1	r_{ag}	r_{bg}	r_{cg}	r_{dg}	r_{eg}	r_{fg}
a	r_{ag}		·72	·63	·54	·45	·36
b	r_{bg}	·72		·56	·48	·40	·32
c	r_{cg}	·63	·56		·42	·35	·28
d	r_{dg}	·54	·48	·42		·30	·24
e	r_{eg}	·45	·40	·35	·30		·20
f	r_{fg}	·36	·32	·28	·24	·20	

r_{ag}, r_{bg}, r_{cg}, r_{dg}, r_{eg} and r_{fg} are the correlations or saturations of

the tests *a. b. c. d. e. f*, with *g*. Let us examine the first two columns:

I	r_{ag}
r_{ag}	
r_{bg}	·72
r_{cg}	·63
r_{dg}	·54
r_{eg}	·45
r_{fg}	·36

Tetrad Differences

We have already noted that in the hierarchical order the correlation coefficients in the columns of the matrix tend to be in the same ratio. Let us take out any group of four coefficients from the matrix

Test	*d*	*e*
a	·54	·48
b	·45	·40

when $\cdot 54 \times \cdot 40 = \cdot 45 \times \cdot 48$

or $\cdot 54 \times \cdot 40 - \cdot 45 \times \cdot 48 = 0$

This is called a *tetrad difference* and this one is

$$r_{ad} \times r_{be} - r_{bd} \times r_{ae} = 0^1$$

Thus, another way of putting Spearman's discovery is that the tetrad differences tend to be zero.

Spearman gives his tetrad equation in the form:

$$r_{ap} \times r_{bg} - r_{ag} \times r_{bp} = 0$$

When this equation holds throughout any table of correlations, and only when it does, every individual measurement of every ability or any other variable contained in the table can be divided into two parts: 'The one part has been called the *general factor*

[1] Those who have some knowledge of determinants will see in this a minor determinant solved by cross-multiplying.

and denoted by the letter g; it is so named because, although varying freely from individual to individual, it remains the same for any one individual in respect of all the correlated abilities. The second part has been called the *specific factor* and denoted by the letter s. It not only varies from individual to individual, but even for any one individual from one ability to another.'[1] (Spearman's two-factor theorem is a piece of general mathematical analysis and is in no way confined to psychology.)

(The precise mathematical expression of the divisibility into two parts is given in the following equation:

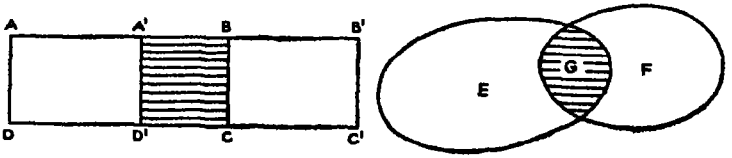
$$m_{ax} = r_{ag} \cdot g_x + r_{as} \cdot s_{ax}.)$$

In this case the sum of the squares of all the scores comes to the number of persons (N). If we take the average by dividing by N we are left with the relationship.

(saturations with $g)^2$ + (saturations with $s)^2 = 1$.

$g^2 + s^2 = 1$ (the 'variance of the test')

communality + specificity = variance



The area of each oval E and F, and each rectangle ABCD and $A^1B^1C^1D^1$ represents the variance of an ability or test The shaded overlap represents the covariance which will equal the correlation coefficient if the areas of each of the rectangles and ovals can be taken as unity Where this is not the case the correlation is given by dividing the area of the overlap by the root of the product of the ovals.

We can now express the tests in the form of equations containing g and s.

e.g. Taking a saturation of g of ·9.

$$\cdot 9^2 + s^2 = 1$$

$$\therefore s = \sqrt{1 - \cdot 81}$$

$$= \sqrt{\cdot 19}$$

$$= \cdot 436$$

[1] *The Abilities of Man*, page 75.

Hence if z is the score of a person in the test given by the suffix to z

$$\begin{aligned} z_a &= .9g + .436\, s_a \\ z_b &= .8g + .600\, s_b \\ z_c &= .7g + .714\, s_c \\ z_d &= .6g + .800\, s_d \\ z_e &= .5g + .866\, s_e \\ z_f &= .4g + .917\, s_f \end{aligned}$$

The six saturations with 'g' are therefore:

$$.9 \quad .8 \quad .7 \quad .6 \quad .5 \quad .4$$

and every correlation coefficient in the matrix can be seen to be the product of two of these saturations e.g.,

$$.56 = .8 \times .7$$
$$\text{or } r_{bc} = r_{cg} \times r_{bg}$$

Now all scores and tests have been standardized, that is they have been given as differences from the average, being given a plus sign if above and a minus sign if below and further have been divided by the standard deviation of each set. The standard deviation of these 'z' scores is therefore unity, and so is the variance of each test (variance = square of standard deviation). Thus the sum of the squares of the saturations of all the 'factors' equals unity (the total *variance*).

We have already seen that the tetrad equation $r_{13}\, r_{24} - r_{23}\, r_{14}$

A Note on Tetrad Relations.

Adapted from Piaggio, *Mathematical Gazette*, Vol XVII, No. 222.

Suppose that we have k sets of numbers denoted briefly by A B . . . and that these are expressible in terms of $(k + 1)$ other sets G, S_A, S_B, . . . no two of which are correlated and $2k$ constants m_a, m_b, . . . n_a, n_b, . . by equations such as:

$$a = m_a g + n_a s_a \quad \ldots (1)$$
$$b = m_b g + n_b s_b \quad \ldots (2)$$

Each equation really denotes N equations as a can take any one of the values, $a_1\, a_2$ . . . with a corresponding set of values for g and s_a. But m_a and n_a are constants which occur unchanged in each of the N equations. Taking the arithmetic mean of the N expressions of a given type (called averaging) gives us:

$$\begin{aligned} &\text{average of } a &&= 0 \\ &\text{average of } a^2 &&= \sigma_a^2 \\ &\text{average of } ab &&= \sigma_a\, \sigma_b\, r_{ab} \end{aligned}$$

$= 0$ is really another way of writing the minor determinant which represents the intercorrelations of two tests with two others.

	3	4
1	r_{13}	r_{14}
2	r_{23}	r_{24}

The process can be extended and tetrad differences of tetrad differences can be found.

Suppose we extend the tetrad (or a minor determinant of order two) to a nonad (or a minor determinant of order three). We could obtain this from the correlation coefficients of three tests 1. 2. 3 with three others 4. 5. 6.

	4	5	6
1	r_{14}	r_{15}	r_{16}
2	r_{24}	r_{25}	r_{26}
3	r_{34}	r_{35}	r_{36}

It is at once evident that this minor determinant of order three can be divided into four determinants of order two (or tetrads):

$$r_{14}\,r_{25} - r_{24}\,r_{15}$$
$$r_{14}\,r_{35} - r_{34}\,r_{15}$$
$$r_{14}\,r_{26} - r_{24}\,r_{16}$$
$$r_{14}\,r_{36} - r_{34}\,r_{16}$$

This is done by taking the top left coefficient r_{14} as the 'pivot'. The four tetrad differences are themselves formed into a tetrad and this can be evaluated. This operation is known as *pivotal*

If all the numbers have been reduced to standard measure (i e., mean of numbers $= 0$ and $\sigma = 1$) these averages reduce to 0, 1 and r_{ab} respectively.

From equations (1) and (2) we get

$$ab = m_a m_b g^2 + m_a n_b g s_b + m_b n_a g s_a + n_a n_b s_a s_b$$

from which by averaging and noting that g and s are uncorrelated

$$r_{ab} = m_a m_b \ldots (3)$$

Similarly $r_{cd} = m_c m_d$ and so on.

Hence $r_{ab}\,r_{cd} - r_{ac}\,r_{bd} = 0$

By permuting the letters a, b, c, d we get three such relations, but only two are independent.

condensation.[1] It must be remembered that the result, if not zero, has to be divided by the product of all the pivots except the last.

If we do not include the numbers in the diagonals which represent the self-correlation of a test, we can reduce the minor determinants of orders two and upwards in the correlation matrix and it may happen that all the minors of a particular order vanish. The 'rank' of the matrix is equal to the order of its greatest non-vanishing matrix (in terms of its rows) and is one less than the orders of the minors which vanish.

Thurstone has shown that a set of tests can be analysed into a number of factors, common to each test, equal to the rank of their correlation matrix plus a specific factor for each test. The factor 'loadings' or 'saturations' in each test can be determined by using the 'centroid' or 'centre of gravity' method. It is called the 'centroid' method because Thurstone conceived it as a means of finding a *centroid* or *centre-of-gravity* in a geometrical model. As we have already seen it is easy to make a model which contains only three vectors (whether these are test-scores or factors) but 4 — or more — dimensional space, though it offers no particular difficulty to the mathematician, cannot be modelled in the ordinary 'Euclidean' way. The geometry of 'hyperspace' is a logical extension of that of three dimensions and it usually yields readily to analytical treatment. That is to say, instead of worrying about the difficulty or impossibility of making useful models we can find and develop the simple algebraic equivalent.[2]

Spearman's work has not gone unchallenged. Although it is true to say that the tetrad differences of Spearman's hierarchies were either zero, or were normally distributed about zero, it must be confessed that there was a tendency to consider too few cases and perhaps to overlook tests which did not fit in with the hierarchy.

[1] See Turnbull and Aitken, *Theory of Canonical Matrices*, or Thomson, *The Factorial Analysis of Human Abilities*, Chapter VI.

[2] 'The student who is not able to work through Thomson's *The Factorial Analysis of Human Ability* or Burt's *The Factors of the Mind* may obtain a simple account of modern work in this field in Thomson's booklet *Some Recent Work in Factorial Analysis* and in Burt's review of Thomson's books in *The British Journal of Educational Psychology*, Vol. XVII, February 1947.

Spearman and his school analysed the results of too few tests, and too readily assumed that all the tetrad differences were normally distributed about zero. Later, many tests were found which did not fit in with the two-factor theory, and group factors had to be admitted. Thurstone of Chicago using a more extended analysis showed that the Spearman results were only a particular case of a larger generalization. It is beyond the scope of this introductory work to give a detailed account of Thurstone's various methods. As in other cases they can be thought of in geometrical and in corresponding algebraical terms. For the purpose of explanation the former method is useful but it is the analytical processes (matrices and determinants) which are actually used for calculating the factors.

Other workers have found group factors, such as a verbal factor *v* which is common to a group of tests but not to all. This could be represented like this:

GROUP FACTORS WITH *g* AND *s*

Test	*General factor*	*Group factors* *a*	*b*	*c*	*Specific factors*
A	x	x			x
B	x	x			x
C	x		x		x
D	x		x		x
E	x		x		x
F	x		x		x
G	x			x	x
H	x			x	x
I	x			x	x

SPEARMAN'S *g* AND *s*

Test	*General factor*	*Specific factors*
A	x	x
B	x	x
C	x	x
D	x	x
E	x	x
F	x	x
G	x	x
H	x	x
I	x	x

As we have already seen, the pioneer work of Spearman described in *The Abilities of Man* with his *g* and *s* factors was limited. Doubtless, he was justified in drawing the conclusions which he arrived at from the mental tests which he applied and the analysis of his results. Nevertheless, further researches have shown the need for more factors and the need for group factors which are common to a limited number of test results. Some method of multiple-factor analysis had to be found to deal with group factors and to obviate the restriction of no correlation except through a factor common to all tests.

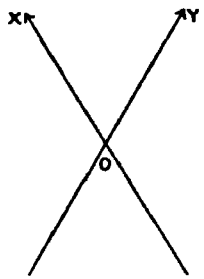
It is beyond the scope of this work to deal with the methods of multiple-factor analysis. There is a considerable literature on the subject and the student would do well to start his study of the matter with Thomson's excellent *Factorial Analysis of Human Ability*. Multiple-factor analysis has been developed by Sir Cyril Burt in England and L. L. Thurstone and H. Hotelling in America.

The most popular method at present in use is that due to Thurstone, or some modification of it. At the time of writing this book the exact nature of the 'factors of the mind' is still a matter of much discussion between psychologists. Even on the cognitive side of mental activity various claims are put forward by different workers concerning the nature, number and importance of these factors. It is too early to decide whether they bear some relation to neurological qualities of the brain, whether they are mathematical artefacts, whether they are just convenient mathematical symbols or whether they represent fundamental quantities in human cognition.[1] (Attempts to submit the affective and conative aspects of mental activity to factorial analysis are fraught with even greater difficulty. The factors suggested by various psychologists, which describe temperament and personality, are legion. Raymond Cattell has listed over 1,000 traits which he has gathered together and arranged in more than fifty 'factors'. It is too early to see whither this will lead us. It will suffice for the student to know that there are well-marked personality traits, such as 'ascendency-submission' which are tested by questions and marked according to a given scale).

A fruitful way of regarding tests, their correlations and factors is to represent them as vectors or straight lines. Two lines may be drawn through a point to represent the tests *and the correlation*

[1] Various leading psychologists in Britain and America have different ways of regarding factors. Thomson and Allport and Anastasi maintain that factors are statistical artefacts without any reality or neurological counterpart. Burt regards them as principles of classification described by selective operators, whereas Spearman originally thought of them as fundamental functions of the mind. Guilford calls them fundamental dimensions of the mind and the Americans Thurstone and Holzinger regard factors as primary or fundamental abilities. The student need not be unduly worried about this. The atomic physicist is up against similar problems when he is considering such problems as the idea of the 'reality' of an electron.

between them is numerically equal to the cosine of the angle made by the two lines. The point of intersection of the lines represents

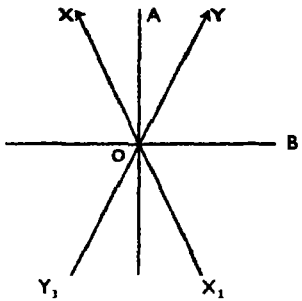


a person who has made an average score on both tests and other points on each line represent standardized scores in the tests the positive direction being shown by the arrows. The degree of correlation increases as the angle decreases and will be perfect positive (+1) correlation when the lines coincide, there will be zero correlation when they are at right angles and negative correlation when the angle becomes obtuse. Any point on the paper represents the scores of a person in each of the tests and each score is given by the perpendicular distance of the point from one of the lines.

The idea of zero correlation when the lines are at right angles (cosine 90° = 0) is a useful one. Sometimes factors can be thought of as vectors which are at right angles. They are then wholly independent factors and have no common quantity or overlap. Instead of speaking of them as rectangular factors we use the Greek ORTHOGONAL to describe them. The factors for which Spearman sought would thus be spoken of as orthogonal. *Obligue factors* are those which could be represented by lines at an angle with one another which is less than a right angle. Most of the methods originated by Alexander, Thurstone and other recent workers use oblique factors.

Let us represent two tests by the lines X^1X and Y^1Y meeting at O. The cosine of angle XOY = the correlation between the tests. A testee with average marks in both tests will be at the point O and other testees will be represented by swarms of dots, like bullet holes

round a bull's-eye O, with the density of dots per unit area becoming smaller the further we go from O. Now the analysis



of test results is equivalent to referring these tests vectors to axes at right angles and these latter will represent *orthogonal* factors. Consider the simplest case of two factor vectors OA and OB respectively bisecting the angles between the test vectors. This was the idea with which Hotelling started his analysis. OA and OB would represent his 'principal components'. There is no necessity, however, for OA and OB to be placed in the position we have taken. They could be placed anywhere provided that they passed through O and were at right angles (orthogonal). These factor vectors can be rotated to the most convenient position, indeed, if either OA or OB are made to coincide with either OX or OY one of the factors is given by one of the test vectors.

When OA bisects the angle XOY, as it does in the case we have given, the scores along OA clearly give the best representation of the results of the two tests. Such a vector is known as the 'first principal component'. (Hotelling.)

In the case of a Spearman analysis of two tests three orthogonal factors would be necessary, that is, a common *g* and two separate *s* factors. Thus his factors may be represented by three straight lines at right angles meeting in a point like three edges of a rectangular box meeting at a corner. These three vectors (still remaining at right angles to one another) are rotated until one is at right angles to the first test and another is at right angles to

the second test. Then, g is represented by the third vector. In general, Spearman's 'two-factor' analysis requires one more dimension in space than the number of tests. Again, we have to use the geometry of 'hyperspace' and models are of only limited help.

If we wish to add a third test to those which we have represented by the two lines through the point O on a plane surface we shall have to consider three-dimensional space. We shall find from trigonometrical tables angles whose cosines are the correlation coefficients of the third test and each of the other two respectively. We shall then find a line going through O which makes these angles with the first two vectors. Usually we shall obtain a kind of tripod with one of the vectors coming out of the plane of the paper. If the sum or the difference of the angles which we have found is exactly equal to the angle between the two original test lines, the three lines will lie in the plane of the paper. Again, if any two angles together are less than a third angle it will be impossible to draw the third line. It will be 'imaginary' in the mathematical sense. More than three tests demand the use of multi-dimensional space and although this cannot be visualized, it is nevertheless a useful mathematical device for work with four or more tests.

Note on Correlation Matrices and Lines of Regression

Consider the following correlation matrix in which x_0, x_1, x_2 . . . etc. are tests of certain aptitudes:

	x_0	x_1	x_2	x_3	. .	x_n
x_0	1	r_{01}	r_{02}	r_{03}	. .	r_{0n}
x_1	r_{01}	1	r_{12}	r_{13}	. .	r_{1n}
x_2	r_{02}	r_{12}	1	r_{23}	. .	r_{2n}
x_3	r_{03}	r_{13}	r_{23}	1	. .	r_{3n}
:	:	:	:	:		
:	:	:	:	:		
x_n	r_{0n}	r_{1n}	r_{2n}	r_{3n}		1

Each of the correlation coefficients r may also be considered as the regression of the score in one test on that of another. In other words, the estimated score in one ability or aptitude is expressed as a linear function of the scores in a number of others x_1 x_2 $x_3 \ldots x_n$. The regression equation becomes:

$$x_0 = b_1x_1 + b_2x_2 + b_3x_3 \ldots + b_nx_n$$

where b_1, b_2, $b_3 \ldots b_n$ are the regression coefficients.

It is sometimes necessary to know how far estimates made from regression equations differ from the true values.

This is given by the multiple correlation (Rm) between the estimates and the true values.

$$\text{Now } \mathrm{R}m = \sqrt{b_1r_{01} + b_2r_{02} + b_3r_{03} + \ldots b_nr_{0n}}$$

Those who have some knowledge of determinants will see that this may be expressed as

$$\mathrm{R}m = \sqrt{1 - \frac{\Delta}{\Delta_{00}}}$$

where Δ is the complete correlation determinant (or matrix) given above and Δ_{00} is the minor determinant which is left when the first row and column are removed.

Similarly we could use the second regression equation and find estimates of x when y is given and these errors of estimate would be distributed with a standard deviation:

$$\sigma_x \sqrt{1 - r^2_{xy}}$$

Here we find again the *alienation* (k) where $k = \sqrt{1 - r^2}$.

CHAPTER VIII

CHI-SQUARED AND CONTINGENCY

ONE of the most useful methods of investigating the numerical results of educational research is the use of chi-squared χ^2. Pearson developed this at the beginning of the present century and in recent years it has become popular in attacking many problems requiring the analysis of variance. The most common and straightforward use of χ^2 is that of testing the agreement between observed quantities and those expected in view of an apparently suitable hypothesis. For instance, we might wish to find whether a set of measures fit a normal distribution curve to such an extent that any discrepancies are due to errors of sampling and are not significant.

If F_e is a number expected and x is the difference between this and the actual number observed F (i.e. the observed number $F = F_e + x$)

then $$\chi^2 = \Sigma \left(\frac{x^2}{F_e}\right)$$

It is obvious that in the case of perfect agreement between the observed and expected values χ^2 will vanish and its value will be smaller in accordance with the closeness of agreement between the sets of values. Tables have been prepared which give a value for P, the proportion of cases in which any value of χ^2 is exceeded. The tables give the relations between χ^2 and P, the probability for various values of n, which must be an integer and represents the number of *degrees of freedom* or *independent variates* of the observed classes. In educational investigations there arise many cases where we might wish to find whether the differences between theoretical or predicted values and those actually observed were due to chance errors of sampling or whether the differences are significant. The chi-square method is also useful to test the 'goodness of fit' of a set of given values to those represented by a standard curve. For example, we know from tables the values of the ordinates of the normal probability curve at various sigma

distances from the mid-point. We may be given a set of values to fit to the curve[1] and the 'goodness of fit' may be estimated by χ^2. Again, we may wish to compare teachers' estimates of pupils' work in classes (A. B. C. D etc.) with their subsequent achievements in examinations. Again, we may wish to compare groupings or estimates with respect to one factor, quality or attainment with those of another. Here we use a contingency table and from this we may obtain a value for the probability that the differences are not due to chance. χ^2 does not normally measure correlation; it is really a measure of divergence rather than association.

Example: The following table gives the theoretical frequencies f_e and the observed frequencies f in fitting values to a normal curve at the given intervals Find whether the fit is good and whether any deviations from normal distributions are due to chance fluctuations.

$$\chi^2 = \frac{\Sigma(f-f_e)^2}{f_e}$$

The table should be set out as follows:

Interval	*Frequencies*		$(f-f_e)$	$(f-f_e)^2$	$\frac{(f-f_e)^2}{f_e}$
	f	f_e			
280-340	17	15	2	4	·27
260-280	13	15	− 2	4	·27
240-260	20	20	0	0	·00
220-240	27	24	3	9	·38
200-220	23	25	− 2	4	·16
180-200	19	21	− 2	4	·19
160-180	15	17	− 2	4	·23
100-160	23	20	3	9	·45
Totals	157	157	0		$\chi^2 = 1{\cdot}95$

Knowing seven of the observed frequencies and the total, we could find the eighth. Thus, there are $(8 - 1) = 7$ degrees of freedom. By consulting the Fisher or Elderton tables for 7

[1] See page 75.

degrees of freedom and $\chi^2 = 1{\cdot}95$ we find a probability value of $P = {\cdot}96$. This means that even if the function were distributed normally throughout all its measures, as great a discrepancy as we have obtained would occur in samples 96 times in 100. The fit is in fact better than usual for the most probable value of P for a true fit is ·50. [If the process were repeated for many samples with the same mean and standard deviation the number of degrees of freedom would be two less, i.e. 5. The value for P in this case would be ·84.]

It often happens that it is necessary to determine the degree of association between two sets of measures which are not normally distributed but are given in the form of numbers in each of a series of classes in both sets of measures. For instance, we may mark a set of Physics papers in four classes A. B. C. and D without further distributions within each class. In the same way we may mark a set of Chemistry papers in four (or some other number of) classes of merit A. B. C. D.

We wish to find whether there is a significant degree of association between the two sets.

It is convenient to arrange the number of cases which fall into each group (the frequency in the group) in a cell in a square or rectangle.

PHYSICS

Chemistry	D	C	B	A	Add
A	1	0	3	6	10
B	2	5	5	1	13
C	3	3	1	2	9
D	4	3	0	1	8
Add	10	11	9	10	40

Total 40

Here we have sixteen cells or categories and each one represents a group in Physics and one in Chemistry so that every possible case is covered. The number in each cell represents the number of students in each category, e.g. 6 students have A marks in Physics and in Chemistry, 3 have a D mark in Chemistry and a C mark in Physics. If there were no correlation between the sets of marks we might expect the 10 students with A.s in Chemistry to be distributed in the proportion 10. 11. 9. 10 in their Physics groups, that is to say, about equal numbers in each group.

Suppose now that there were no relationship between the groups in Chemistry and those in Physics. Let us calculate how many students would fall into each of the 16 cells in this case. (F_e is the expected frequency.)

F_e for A in Chemistry and D in Physics $= \frac{10 \times 10}{40}$

F_e for A in Chemistry and C in Physics $= \frac{10 \times 11}{40}$

F_e for A in Chemistry and B in Physics $= \frac{10 \times 9}{40}$

and so on.

Now make a 4 × 4 table of these F_e.s:

	D	C	B	A	
A	2·50	2·75	2·25	2·50	
B	3·25	3·57	2·92	3·25	
C	2·25	2·47	2·02	2·25	
D	2·00	2·20	1·80	2·00	

TABLE OF F_e.S

	D	C	B	A	
A	1·5	2·75	·75	3·50	
B	1·25	1·43	2·08	2·25	
C	·75	·53	1·02	·25	
D	2·00	·80	1·80	1·00	

TABLE OF $(F - F_e)$

F = actual frequency

Note that in view of later squaring the signs are all written as positive.

The next table gives $\frac{(F - F_e)^2}{F_e}$, that is, the numbers in the last table were squared and divided by their respective F_e.s.

	D	C	B	A	
A	·90	2·75	·25	4·90	
B	·48	·57	1·48	1·85	
C	·56	·12	·50	·03	
D	2·00	·29	1·80	·50	

TABLE OF $\frac{(F - F_e)^2}{F_e}$

The sum of all the $\frac{(F - F_e)^2}{F_e}$ numbers,

i.e. $\frac{\Sigma(F - F_e)^2}{F_e} = \chi^2$ (chi-squared) $= 18{\cdot}98$.

On consulting Fisher's or Elderton's tables the value of P, the probability for $\chi^2 = 18{\cdot}98$ and 9 degrees of freedom[1] is equal to ·025. Thus the chances are 1 in 40 that the deviations of the actual from the expected frequencies could be through chance errors of sampling. Accordingly, we have grounds for believing that there is a contingency or relationship between the variables.

The Coefficient of Mean Square Contingency

The coefficient of mean square contingency is given by

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

In the example we have worked out

$$C = \sqrt{\frac{18{\cdot}98}{40 + 18{\cdot}98}} = {\cdot}57$$

Contingency is a better measure of divergence than association and should be regarded as such. Nevertheless, if the number of cells used were increased and a finer grouping obtained, C would approach in value to that of the correlation only *if* the distributions of both sets of measures were normal or nearly normal.

A Note on Degrees of Freedom

Chi-squared tables give the value of the probability P in terms of χ^2 and the number of degrees of freedom. This number is not usually equal to the number of cells in the contingency table or the number of cases, but is usually one less. Nevertheless, as R. A. Fisher has shown, the number of degrees of freedom, when the marginal totals remain the same sample after sample, will be $(c - 1)(r - 1)$ where c is the number of columns and r is the

[1] See below.

number of rows. We have to ask ourselves how many cells could be filled in from prior knowledge and subtract this from the total number of cells in order to obtain the number of degrees of freedom; e.g. if we have a 4×4 table and can assume that the marginal totals remain fixed we should be able to compute the fourth row or column in each case knowing the three others.

The number of degrees of freedom is therefore

$$(4 - 1)(4 - 1) = 9.$$

Note on Student's 't'[1]

Student's 't' is defined as

$$t = \frac{x}{\sigma_x}$$

where x is the deviation of a measure from the true value which is assumed from a normal distribution and σ_x is the standard deviation of all the measures in the sample. Student worked out the distribution of t (which he originally called z) and found that it was particularly useful for working with small samples. At first Student carried his table only to $N = 10$ and found that the standard error of his distribution was $\frac{1}{\sqrt{N - 3}}$ and later Fisher developed the table in terms of $N - 1$ degrees of freedom. Most of Fisher's tables are constructed so that a probability of 5% (odds of 20 to 1) is significant and a probability of 1% is highly significant. In the case of a normal distribution (n very large) probability of 5% corresponds to a t of 1·96 and a probability of 1% corresponds to a t of 2·58.

[1] 'Student', whose real name was William Sealy Gosset, died in 1937. He was a senior member of the brewing firm of Guinness in whose service he developed much of his statistical work. He chose his pseudonym out of respect for the 'master' Karl Pearson.

CHAPTER IX

THE ANALYSIS OF VARIANCE

STANDARD deviation has proved so useful as a measure of dispersion, as a step to correlation, factor analysis and the use of the normal curve that the more recent and often more useful technique of the analysis of variance has tended to be overlooked. It is possible that the influence of Spearman, who made such great use of correlation coefficient in his technique of factor analysis did something to hinder the development of the more widespread use of the analysis of variance.[1]

Variance may be regarded as the square of the standard deviation

$$\text{If } \sigma = \sqrt{\frac{\Sigma d^2}{N}}$$

$$V = \frac{\Sigma d^2}{N}$$

where N is the number of measures and d is the deviation of a measure from the mean of all the measures.

(If the measures have been standardized by arranging them as deviations from their mean and dividing them by the standard deviation the S.D. is therefore the unit of measurement, i.e. S.D. = 1 and V = 1.)

If we regard the mean as the first moment about the point from

[1] As has already been noted the psychologist of a generation ago borrowed something of the terminology and technique of the Galton-Pearson school of biometricians. In recent times the work of Professor R. A. Fisher, formerly of the Rothamsted Experimental Station, in statistics chiefly concerned with agriculture and other biological investigations has been adapted to psychological needs, particularly by Sir Cyril Burt in this country. The most valuable aspects of Fisher's work for our purposes are (*a*) his methods of designing experiments so that the results shall be susceptible to simple statistical treatment (*b*) the analysis of variance. Details of his methods (with particular reference to agriculture) are to be found in *Statistical Methods for Research Workers* and *Design of Experiments*. Burt's expositions have a simplicity and clarity not always to be found in these treatises.

which the mean is measured the variance of a distribution may be defined as the second moment about the mean:

$$V = \frac{1}{N}\Sigma(x - \bar{x})^2$$

where x is a score or measure
and $\bar{x}$ is the mean of the whole distribution.

Variance as a measure of variability has an advantage because it is *additive*, that is, the total variance of a set of measurements may be regarded as the sum of the independent parts or 'factors' which combine to make up the variance.

$$\sigma_x^2 = \sigma_a^2 + \sigma_b^2 + \sigma_c^2 + \ldots \text{ etc.}$$
$$\text{if } x = a + b + c.$$

In the analysis of variance the process is reversed and the total variance is broken down into those of the several components. One of these variances will obviously be due to error in measurement and usually will be taken to consist of random errors due to the smallness of the size of the sample which has been used for the investigation. *The most frequent and useful application of the analysis of variance is to compare the significance of the variance due to some particular factor with the amount of variance due to error.*

(It will be recalled that in factor analysis the factors have to be discovered in the process of the analysis and their relative amounts estimated. In the analysis of variance the possible factors are assumed by reference to the given data and the problem is to establish their relative significance, that is, to find what is the probability that the variance due to each factor is to be accounted for as an effect of pure chance. In factor analysis we try to determine the relative importance of the inferred factors.)

Let us consider a set of marks (x) which have been correlated with another set (y). Were all the individuals in the x column to have the same value there would still remain some scatter in the y column, that is, when x is constant there is yet some variability in the y scores. When there is correlation between the x and y values the variability expressed as a ratio is $\frac{\sigma_c^2}{\sigma_y^2}$. As this is the

proportion of the variance (σ^2) remaining when x is constant it may be considered the proportion of the variance in y attributable to factors in y other than x. Conversely, the reduction in variance when x is kept constant is the part of the variance due to x factor. In terms of the entire variance of y the ratio is

$$\frac{\sigma_y^2 - \sigma_c^2}{\sigma_y^2} = 1 - \frac{\sigma_c^2}{\sigma_y^2}$$

But $r = \sqrt{1 - \frac{\sigma_c^2}{\sigma_y^2}}$ Therefore $r^2 = 1 - \frac{\sigma_c^2}{\sigma_y^2} = \frac{\sigma_y^2 - \sigma_c^2}{\sigma_y^2}$

Accordingly the total variance may be divided into two parts of which the proportion due to what is common to x and y is equal to r^2, and the proportion due to the other factors is

$$\frac{\sigma_c^2}{\sigma_y^2} = 1 - r^2$$

r^2 is known as the *coefficient of determination.*

[The above is true when correlation is linear and the line of regression is straight. Nevertheless, a similar relationship exists when the correlation is not linear and the correlation ratio η (eta) is used. In this case, the proportion of variance of y is separable into two parts: that due to x is $\frac{\sigma_m^2}{\sigma_y^2} = \eta^2$ and that due to the other factors $\frac{\sigma_c^2}{\sigma_y^2} = 1 - \eta^2$.]

In the analysis of variance the easiest way is to consider the average for each class implied by the factor. As, for example, we might require to find whether on the average males or females are more intelligent. All we have to do is to find the respective means of intelligence-test scores and to determine whether the difference between the two means can be attributed to the effects of random sampling. Here the classification is dichotomous but if we have to consider, in addition to sex, differences arising from race or school, we should have multiple classification and should have to compare a number of means all derived from the same principle of classification.

Thus, it is useful in the case of the simple sex classification to find the standard error of the difference between the two averages, for this will tell us whether the difference is significant or attributable to chance errors of sampling.

$$\text{S.E. of a difference of means} = \sigma_D = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

N_1 and N_2 are the numbers in each of the two sets respectively and σ_1 and σ_2 are their standard deviations:

$$\text{P.E.} = .6745\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

The S.E. divided into the difference between the averages should give a quotient of at least 3, though if it were above 2 it might be worth while continuing the investigation.

Other standard errors which are useful in educational research are as follows:

Standard error of a difference between the averages of scores which are intercorrelated. If we wish to consider the significance of the difference between the averages of scores in two tests or in repeated tests taken by a single set of persons

$$\text{S.E.} = \sigma_D = \sqrt{\frac{1}{N}(\sigma_1^2 + \sigma_2^2 - 2r\sigma_1\sigma_2)}$$

or if σ_1 and σ_2 are taken to represent the S.E.s of the *means* of the original scores and *not* the S.D.s of the original scores

$$\sigma_D = \sqrt{\sigma_1^2 + \sigma_2^2 - 2r\sigma_1\sigma_2}$$

In view of the differences which arise through errors of sampling the average of a sample may vary from the true average which would be found if we were able to take a very large number.

$$\text{The S.E. of the mean or average } \sigma_M = \frac{\sigma}{\sqrt{N}}$$

where σ is the standard deviation of the original sample.

In the same way differences in the nature of samples ('errors of sampling') may cause errors in the S.D.s of a sample.

The standard error of a standard deviation $\sigma_\sigma = \frac{\sigma}{\sqrt{2N}}$

The standard error of a difference between two standard deviations is equal to

$$\sqrt{\frac{\sigma_1^2}{2N_1} + \frac{\sigma_2^2}{2N_2}}$$

where σ_1 and σ_2 are the standard deviations of N_1 and N_2 are the number of cases in the respective groups or sets

Standard error of a percentage and of a difference between percentages.

If x is the percentage then

$$\text{Standard Error of } x = \sqrt{\frac{x(100 - x)}{N}} = \sqrt{\frac{100x - x^2}{N}}$$

and the standard error of a difference between two percentages x_1 and x_2 is

$$\sqrt{\frac{x_1(100 - x_1)}{N_1} + \frac{x_2(100 - x_2)}{N_2}}$$

The formulae are most useful in finding the numbers of cases which it is necessary to investigate in order to be certain that percentage differences between groups are significant, e.g. It appears from dental records that 40% girls and 43% boys at certain schools are in need of dental treatment. What is the minimum of children which we must take in order to make sure that the 3% difference is significant?

If the difference of 3% is reliable it should be more than 3 times its S.E.

$\therefore$ S.E. should not be greater than 1%

$$\therefore 1 = \sqrt{\frac{40 \times 60}{N} + \frac{43 \times 57}{N}}$$

$$\therefore N = 4851$$

Thus to make sure that the 3% difference is significant the investigation should be based on the examination of 4851 (say 5000) boys and an equal number of girls.

Problem

A test has been applied to five arts students and five science students. The marks obtained are given below. The average for the arts students is 3 marks more than that of the science students. With this small sample is this difference likely to be a matter of chance or is it safe to assume that arts students are better on the average?

	Arts Students				*Science Students*		
Name	*Marks*	*Deviation*	*Square*	*Name*	*Marks*	*Deviation*	*Square*
Cowper	21	+ 1	1	Maxwell	19	+ 2	4
Shaw	19	− 1	1	Faraday	14	− 3	9
Scott	18	− 2	4	Darwin	18	+ 1	1
Stewart	23	+ 3	9	Dale	15	− 2	4
Lamb	19	− 1	1	Newton	19	+ 2	4
Totals	5)100	0	16		5)85	0	22
Mean	20			Mean	17		

$$\text{Average of means } \frac{20 + 17}{2} = 18.5$$

Deviation + 1·5 Deviation − 1·5

To obtain the standard deviation we divide not by the number of each set of cases but by the number of *degrees of freedom*. This is an important conception in statistical analysis. In each column there are 5 deviations from a mean calculated from the given data. But the total of all the 5 deviations must be zero, and thus if we know 4 deviations we can at once calculate the 5th. Accordingly there are 4 degrees of freedom, i.e. only 4 deviations are independent.

Thus the standard deviation of the individuals in the sample is

$$\sigma = \sqrt{\frac{\Sigma x^2}{n_1 + n_2 - 2}} = \sqrt{\frac{16 + 22}{8}} = \sqrt{\frac{38}{8}}$$
$$= \sqrt{4{\cdot}75} = 2{\cdot}179$$

and the standard deviation of the difference is

$$\sigma_d = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 2{\cdot}179 \sqrt{\frac{1}{5} + \frac{1}{5}} = 1{\cdot}376$$

The critical ratio t is given by

$$\frac{\text{mean}_1 - \text{mean}_2}{\sigma_d} = \frac{20 - 17}{1{\cdot}376} = \frac{3}{1{\cdot}376} = 2{\cdot}176$$

On consulting Yule and Kendall's 't-table' we find that for 8 degrees of freedom the probability of obtaining a difference as large as this is P = 2(1 − ·97) = ·06 or 6%. The probability of getting a difference as large as this by chance is 6 to 100, that is, the odds against getting a difference as large as this by chance are about 15 to 1. The difference cannot therefore be accepted as really significant.

Instead of comparing the difference between the means with a standard deviation derived from the individual measurements we can compare the variance of the means with a variance based on the original measurements.

Firstly, let us reduce all the given marks to deviations about the *general* mean. This is

$$\frac{100 + 85}{10} = 18{\cdot}5$$

Then deviation of Art Students mean from General Mean = + 1·5
" " Science " " " " = − 1·5

Now split the marks for each student into three components:

(1) the general average; (2) the deviation of his group mean; (3) his individual deviation above or below the sum of the two means.

Thus Cowper's mark is 21 = 18·5 + 1·5 + 1·0.

MARKS ANALYSED IN DEVIATIONS OF MEANS AND INDIVIDUALS

Deviations of Means		*Deviations of Individuals*		*Total Deviation from General Mean*	
1*a*	1*b*	2*a*	2*b*	3*a*	3*b*
Arts	*Science*	*Arts*	*Science*	*Arts*	*Science*
+ 1·5	− 1·5	− 1·0	+ 2·0	+ 2·5	+ 0·5
+ 1·5	− 1·5	− 1·0	− 3·0	+ 0·5	− 4·5
+ 1·5	− 1·5	− 2·0	+ 1·0	− 0·5	− 0·5
+ 1·5	− 1·5	+ 3·0	− 2·0	+ 4·5	− 3·5
+ 1·5	− 1·5	− 1·0	+ 2·0	+ 0·5	+ 0·5

SQUARES OF THE ABOVE

2·25	2·25	1·00	4·00	6·25	0·25
2·25	2·25	1·00	9·00	0·25	20·25
2·25	2·25	4·00	1·00	0·25	0·25
2·25	2·25	9·00	4·00	20·25	12·25
2·25	2·25	1·00	4·00	0·25	0·25
11·25	11·25	16·00	22·00	27·25	33·25
Total 22·50		38·00		60·50	

CALCULATION OF MEAN SQUARES

Source of Variation	*Degrees of Freedom*	*Sums of Squares*	*Mean Square*
Between Groups	2 − 1 = 1	22·50	22·50
Within Groups	10 − 2 = 8	38·00	4·75
Total	10 − 1 = 9	60·50	(6·72)

VARIANCE-RATIOS, OBSERVED AND EXPECTED

Observed	*Degrees of Freedom*	*Expected*
$F = \frac{22{\cdot}50}{4{\cdot}75} = 4{\cdot}737$	1 and 8	5·32

The deviation of the mean and the deviation of the individual are given in columns 1*a*. 1*b* and 2*a*. 2*b* respectively. It will be seen that these add up to the deviation about the general mean given in columns 3*a* and 3*b*. Further, in the following table it will be seen that the totals of the squares of mean and of individual deviations add up to the total of the squares of the deviation from the general mean.

To obtain the 'mean-squares' or 'variances' we divide each of the three square sums by the corresponding degrees of freedom. There are 2 deviations for the 2 means, but as these are calculated from the general mean of the data one degree of freedom has been lost. There are 5 deviations about the mean for arts students and 5 about the mean for science students, and each set of these is calculated from the mean of its group. Hence the number of degrees of freedom is $(5 - 1 + 5 - 1) = (10 - 2) = 8$. As there are 10 individual deviations about the general mean these give $(10 - 1) = 9$ degrees of freedom.

In the table showing the variance or mean square note that the column of degrees of freedom adds up to the degrees of freedom of the whole group, and the square sums for the two components add up to the square sum of the entire group and this provides a useful check.

As we analyse the total sum of the variances and not the total variance, the variances do *not* add up to the total variance. We now proceed to test the variance between the means of the two groups. (If the variance to be tested is due solely to error, then it should be equal to the error-variance. Hence to test the former we divide by the latter.) The variance of the individuals within the group, taken from the mean of either group, is treated as denoting the error variance. The probabilities corresponding to various values of the error variance F can be found in Fisher's or Snedecor's tables, and as before a 5% probability may be taken as marking the borderline for significance. The table gives 4·737 in this case which is less than the borderline value. Again, by this method we conclude that the difference between the two means cannot be regarded as fully significant.

In the case under consideration $F = t^2$ (and we note that

$\sqrt{F} = \sqrt{4{\cdot}737} = 2{\cdot}176$ which was the value previously obtained for t).

Testing the Significance of the Differences between Several Means[1]

Where the criterion of classification gives two classes only it is adequate to test the difference between the two means by the standard error of the difference, that is, by the t-ratio. When we have three or more classes it is necessary to use methods involving the variance or F-ratio. Suppose that instead of considering the abilities of students in only two faculties of a university, we have to make a comparison of students in all the faculties. Suppose, for simplicity, we consider three faculties only and that the test results are as follows:

MARKS FOR ARTS, SCIENCE AND MEDICAL STUDENTS

	Arts *Mark*	*Science* *Mark*	*Medicine* *Mark*	*Dev.*	*Square*
	21	19	18	+ 2	4
	19	14	16	0	0
	18	18	15	− 1	1
	23	15	17	+ 1	1
	19	19	14	− 2	4
Total	5)100	5)85	5)80		10
Average	20	17	16		
Deviation	+ 2·3	− 0·6	− 1·6		
Square	5·4	0·4	2·7		

It is unnecessary to repeat the deviations and squares for arts and science students. It is also unnecessary to repeat the means, etc., for every person tested. We have simply to multiply the square of each mean by 5 (the number of individuals) and then take the sum; or more simply to sum the squares first ($5{\cdot}4 + 0{\cdot}4 + 2{\cdot}7 = 8{\cdot}6$) and then multiply the sum by 5. We obtain $5 \times 8{\cdot}6 = 43{\cdot}3$.

[1] I am indebted to Sir Cyril Burt for the treatment of this problem and for the subsequent account, taken from his laboratory notes, of his adaptation of Fisher's methods.

The sums of the squares of the individual deviations within each of the three groups (calculating from the corresponding group mean) are $16 + 22 + 10 = 48$. The square-sums for the 15 deviations for the general mean (17·6) need not be calculated, except as a check.

Tabulating the results as before, we obtain the mean squares as follows:

CALCULATION OF MEAN SQUARES

Source of Variation	*Degree of Freedom*	*Sum of Squares*	*Mean Squares*
Between Groups	$3 - 1 = 2$	43·3	21·6
Within Groups	$15 - 3 = 12$	48·0	4·00
Total	$15 - 1 = 14$	91·3	

VARIANCE RATIOS, OBSERVED AND EXPECTED

Observed	*Degree of Freedom*	*Expected*
$F = \frac{21·6}{4} = 5·42$	2 and 12	3·88

The ratio of the two variances is now 5·42, well above the value we should expect with 2 and 12 degrees of freedom. Thus there can now be little doubt that the difference of faculty does after all tend to produce slight but genuine differences in the average marks obtained by the test.

For purposes of illustration we have taken tiny samples with 5 individuals in each. But the numbers in each sample need not be the same, and indeed may be so large that the sums of squares are best calculated from grouped frequencies. With continuous variates it is then better not to use Sheppard's correction but to keep the grouping fine.

The method may be conveniently used to test the significance of the correlation ratio. Treating the groups as 'arrays' in a correlation-table, we have

$$\eta^2 = \frac{\text{Sum of Squares between Groups}}{\text{Total Sum of Squares}} = \frac{43·3}{81·3} = ·533 .$$

Hence $\eta = ·730$ (by consulting Yule and Kendall table, p. 454).

Two Criteria of Classification

Testing the Significance of a Difference between two *Means*

Problem: Consider the marks allotted to the four pupils as follows:

	Tom	*Dick*	*Harry*	*George*	*Total*	*Average*
Arithmetic	29	24	14	1	68	17
English	29	28	15	4	76	19
Drawing	32	27	27	22	108	27
Handwork	34	29	28	25	116	29
Total	124	108	84	52	368	92
Average	31	27	21	13	92	23

Take, to begin with, two pupils only. The average mark allotted to Tom is 31, to Dick 27. Can we safely infer from this that Tom's general ability is significantly greater than Dick's, or (since we have used only 4 tests) is it more likely that the difference results solely from chance?

1st Method: Standard Error of the Mean Difference

As before, the most obvious procedure is to calculate the standard error of the difference by the usual formula.

CALCULATION OF STANDARD ERROR OF DIFFERENCE

	1	2	3	4	5
Test	*Tom*	*Dick*	*Diff.*	*Dev.*	*Squares*
Arithmetic	29	24	+ 5	+ 1	1
English	29	28	+ 1	− 3	9
Drawing	32	27	+ 5	+ 1	1
Handwork	34	29	+ 5	+ 1	1
Total	124	108	+ 16	0	12
Average	31	27	+ 4	0	

Since Tom's and Dick's marks may be correlated, it is simpler to calculate the detailed differences instead of the S.D.s of the marks observed and their correlation. The calculation is shown in the first 3 columns of the last table.

The deviations of the differences about the mean difference (+ 4) are given in column 5. As usual, to find their standard

deviation we add the squares of the deviations (column 4), but we divide by the number of degrees of freedom. When we started there were $n = 4$ items, and therefore 4 'degrees of freedom' (i.e. 4 figures that vary independently). But in taking deviations about a mean calculated from the observed data, we have lost one degree of freedom: for, when we know the first 3 deviations (or any 3 deviations), we can fill in the 4th from the fact that the total *must* be 0.

Hence to find the 'mean square' we divide, not by 4 but by 3. This mean square ($12 \div 3 = 4$) is the 'variance' of the individual differences: and its square root (2) would be their standard deviation.

But we require the standard deviation of the mean difference. To obtain the variance of a mean, we divide the variance of the individuals by the number of individuals. We then obtain $4 \div 4 = 1$. The square root of this gives the standard deviation. In the absence of any other information we must take the standard *deviation* of the mean difference thus calculated, as the best indication of the standard *error* of the mean difference. Accordingly, to test the significance of the mean difference (m) we divide it by its standard deviation. Using the t-ratio as before, we obtain

$$t = \frac{\bar{m}}{\sigma_m} = \frac{4}{1} = 4$$

(where $\sigma_m = \sqrt{\{\Sigma x^2 \div n(n-1)\}}$).

From the t-table given by Yule and Kendall (p. 536) we find that, with 3 degrees of freedom, a value of $t = 4$ gives $y = .986$. Thus, the chance of getting a difference so large as this (in either direction) would be $P = 2(1 - .986) = .028$ or 35 to 1 against.

The method indicated above has certain limitations although it suffices for the actual problem which is given. We may desire to test the significance of differences not only between two pupils but between all the pupils in the class, but it would involve a great deal of work to prepare every pair of pupils by the method given. Even if we did this the general picture would still not be clear, as it is impossible to draw the general inference from the pairs

considered severally. We need a more comprehensive analysis of all the data which has been given. This is given by a general method of analysis of variance on the following lines:

The 8 observed marks set out in columns 1 and 2 are formed by the deviations of 8 performances about the average performance of both boys in all four tests (i.e. about the average mark of 29). The 8 deviations are given in columns 3 and 4. Instead of measuring the *total* amount of deviation by the sum of the 8 deviations (which would be zero unless we ignore the signs) we can measure it by the sum of the squares of those deviations. The squares are given in columns 7 and 8.

CALCULATION OF TOTAL VARIANCE FOR TWO BOYS

	Mean		*Deviations*		*Squares of Means*		*Squares of Deviations*	
	Tom	*Dick*	*Tom*	*Dick*				
Test	1	2	3	4	5	6	7	8
Arithmetic	29	29	0	− 5	841	841	0	25
English	29	29	0	− 1	841	841	0	1
Drawing	29	29	3	− 2	841	841	9	4
Handwork	29	29	5	0	841	841	25	0
Total			0		3364	3364	34	30
					6728		64	
					6792			

MEANS AND DEVIATIONS

	Means		*Means of Boys*		*Means of Tests*		*Deviations*		*Totals*	
	Tom	*Dick*	*Tom*	*Dick*	*Tom*	*Dick*	*Tom*	*Dick*	*Tom*	*Dick*
	1*a*	1*b*	2*a*	2*b*	3*a*	3*b*	4*a*	4*b*	5*a*	5*b*
Arithmetic	29	29	+ 2	− 2	− 2.5	− 2.5	+ 0.5	− 0.5	29	24
English	29	29	+ 2	− 2	− 0.5	− 0.5	− 1.5	+ 1.5	29	28
Drawing	29	29	+ 2	− 2	+ 0.5	+ 0.5	+ 0.5	− 0.5	32	27
Handwork	29	29	+ 2	− 2	+ 2.5	+ 2.5	+ 0.5	− 0.5	34	29

SQUARES OF ABOVE

Test	1*a*	1*b*	2*a*	2*b*	3*a*	3*b*	4*a*	4*b*	5*a*	5*b*
Arithmetic	841	841	4	4	6.25	6.25	0.25	0.25	841	576
English	841	841	4	4	0.25	0.25	2.25	2.25	841	784
Drawing	841	841	4	4	0.25	0.25	0.25	0.25	1024	729
Handwork	841	841	4	4	6.25	6.25	0.25	0.25	1150	841
Total	3364	3364	16	16	13.00	13.00	3.00	3.00	3862	2930
	6728		32		26		6		6792	

Components

Our task is now to analyse these gross deviations into their chief components. Each deviation may be regarded as the sum of 3 deviations: (i) the mean deviation of the particular boy above or below the general mean (29); (ii) the mean deviation of the particular test above or below the general mean; (iii) the individual deviation of each mark from the sum of these two means. This subdivision is shown in the table of Means and Deviations. Observe that, in combination with the general mean, the three figures add up to the original marks, appended in the last two columns.

We now square all these figures and enter them in the Table of Squares where they are analysed. We notice that the component sums at the bottom of the table add to the grand total (6792).

We are not concerned with the squares of the general mean (6728). What interests us is the partition of the sum of the square of the unanalysed deviations (64) into the sum of the sums of the squares of the three components. We observe that

$$64 = 32 + 26 + 6$$

The Variances

We can now proceed to test the significance, not only of the variance due to the differing means of the 2 boys, but also of the variance due to the differing means of the 4 subjects. As before, what we shall test is not the differences between the means, but the total variance of the means. The sums of squares and the degrees of freedom by which we divide them are tabulated in the first two columns of the table. The result of the division is given in the last column.

ANALYSIS OF VARIANCE: (TWO BOYS)

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*
Boys	32	$2 - 1 = 1$	32
Tests	26	$4 - 1 = 3$	8·6
Error	6	$4 - 1 = 3$	2
Total	64	$8 - 1 = 7$	(9·14)

Degrees of Freedom

Since the deviations of the 2 boys' means and the deviations of the 4 test means are calculated about the general mean, we must deduct one degree of freedom from each. The same is true of the deviations of the 8 performances: but this we only need as a check. The boys' variance and the test variance are the variances to be tested, and so form the numerator of the variance-ratio. And since a variance, not a difference, is being tested, we require for the denominator, not the standard error, but the error variance. The only part of the data that we can use to indicate the error variance will be the deviations of the 8 performances from the sum of the means, i.e. the deviations shown in columns 4*a* and 4*b*. There are 8 figures; but in calculating these figures from the original 8 marks we have already used 5 degrees of freedom (1 for the general mean; 2 − 1 = 1 for the boys' means; and 4 − 1 = 3 for the test means). Hence only 3 degrees of freedom are left. It is easy to see that, if we take any 3 figures in columns 4*a* and 4*b* say + 0·5, − 1·5, + 0·5, we can deduce the other 5, because we know that the sums of both columns and rows must all be zero.

Significance Test

To test significance, we now take the ratios of the variance of the boys' means, and then of the test means, to the variance due to 'error'.

VARIANCE RATIOS (F), OBSERVED AND EXPECTED

Source	*Ratio*	*Observed*	*Degrees of Freedom*	*Expected*
Boys	F_b	$\frac{32}{2} = 16$	1 and 3	10·1
Tests	F_t	$\frac{8.6}{2} = 4.3$	3 and 3	9·3

Thus the difference between the two boys is fully significant, but the differences between the tests (applied to only two pupils in this part of the inquiry) is not significant.

Relation between the Two Alternative Methods

Since, with 1 and 3 degrees of freedom, an F-ratio of 10·1 represents P = 0·05, we might guess, by rough interpolation, that an F-ratio of 16 would represent P = 0·03 or thereabouts (the value obtained with the first method). In fact, we note as before that $F = t^2$, for $F = 4$ and $t = 2$.

Testing Reliability

There is no reason why the two columns of observed figures, like those set out in columns 1 and 2 above, should always represent persons, or the rows should always represent tests. For example, if we had applied two *tests* to four (or more) *persons*, then the headings 'Tom' and 'Dick' would be altered to '1st Test', '2nd Test'; and the side-titles would be the names of the persons tested instead of names of school subjects. This is the form the data take when we wish to test the reliability of two successive applications of the same test. The two means of the *columns* will now represent difficulty of tests, or possibly the improvement shown in the second test as a result of practice or familiarity with the first; and the means of the pair of marks in each row the average ability of the boys tested. Unless the averages for the boys differ significantly, the test is failing to differentiate between the several tested, and so is devoid of reliability. The usual measure of the *amount* of reliability is, of course, the correlation between the two columns.

Testing the Significance of the Differences between SEVERAL *Means*

Problem

The advantages of the second procedure are most evident where we desire to test the significance of the differences between the means, not for two boys only, but for several — say four. As before we can at the same time test the significance of the differences between the means for the four school subjects. Subtracting the general mean (23) from the figures in the table for four boys we have

	DEVIATIONS						SQUARES OF DEVIATIONS				
Test	*Tom*	*Dick*	*Harry*	*George*	*Total*	*Mean*	*Tom*	*Dick*	*Harry*	*George*	*Total*
Arithmetic	+ 6	+ 1	− 9	− 22	− 24	− 6	36	1	81	484	602
English	+ 6	+ 5	− 8	− 19	− 16	− 4	36	25	64	361	486
Drawing	+ 9	+ 4	+ 4	− 1	+ 16	+ 4	81	16	16	1	114
Handwork	+ 11	+ 6	+ 5	+ 2	+ 24	+ 6	121	36	25	4	186
Total	+ 32	+ 16	− 8	− 40	0	0	274	78	186	850	1388
Mean	+ 8	+ 4	− 2	− 10	0	0					

Components

We now analyse these deviations into the same three components as before, namely (i) the mean deviation of each boy; (ii) the mean deviation of each test, (iii) the deviation of each of the 8 performances from the sum of the two means. These are shown in the first table below. The reader should check the fact that for each performance the three components add up to the deviation shown above.

The squares of these deviations follow:

	ANALYSIS OF DEVIATIONS				SQUARES				
	Tom	*Dick*	*Harry*	*George*	*Tom*	*Dick*	*Harry*	*George*	
	(1) *Deviations for Boys*				(1) *Squares of Deviations*				*Total*
Arithmetic	+ 8	+ 4	− 2	− 10	64	16	4	100	184
English	+ 8	+ 4	− 2	− 10	64	16	4	100	184
Drawing	+ 8	+ 4	− 2	− 10	64	16	4	100	184
Handwork	+ 8	+ 4	− 2	− 10	64	16	4	100	184
Square Sum					256	64	16	400	736
	(2) *Deviations for Tests*				(2) *Squares of Deviations*				*Total*
Arithmetic	− 6	− 6	− 6	− 6	36	36	36	36	144
English	− 4	− 4	− 4	− 4	16	16	16	16	64
Drawing	+ 4	+ 4	+ 4	+ 4	16	16	16	16	64
Handwork	+ 6	+ 6	+ 6	+ 6	36	36	36	36	144
Square Sum					104	104	104	104	416
	(3) *Deviations for Performances*				(3) *Squares of Deviations*				*Total*
Arithmetic	+ 4	+ 3	− 1	− 6	16	9	1	36	62
English	+ 2	+ 5	− 2	− 5	4	25	4	25	58
Drawing	− 3	− 4	+ 2	+ 5	9	16	4	25	54
Handwork	− 3	− 4	+ 1	+ 6	9	16	1	36	62
Square Sum					38	66	10	122	236

Error

Provisionally we shall treat the four tests as random (and therefore uncorrelated) specimens of tests for 'general ability': that would mean that we can take the last set of deviations (the residuals) as due to 'error'. Strictly this assumption should be tested first of all: and in fact we shall presently see that it is *not* tenable. But for the present we are concerned only to illustrate the procedure for simple cases first.

Degrees of Freedom

The degrees of freedom are calculated as before. The easiest way to decide the degrees of freedom for the 'error variance' is to subtract from the total degrees (15) the degrees for the other two items ($3 + 3 = 6$): that is equivalent to subtracting from the total *number* (16) the number of constants used to calculate the deviations for error ($1 + 3 + 3 = 7$).

We can now tabulate the calculations for the mean squares (or 'variances') in the same way as before.

ANALYSIS OF VARIANCE: (FOUR BOYS)

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*
Boys	736	$4 - 1 = 3$	$245.\dot{3}$
Tests	416	$4 - 1 = 3$	$138.\dot{6}$
Residual	236	$16 - 7 = 9$	26·2
Total	1388	$16 - 1 = 15$	(92·5)

Significance Test

The variance ratios are calculated as before.

VARIANCE RATIOS (F), OBSERVED AND EXPECTED

Source	*Ratio*	*Observed*	*Degrees of Freedom*	*Expected*
Boys	F_b	$\frac{245.\dot{3}}{26.2} = 9.36$	3 and 9	8.8
Tests	F_t	$\frac{138.\dot{6}}{26.2} = 5.28$	3 and 9	8.8

The degrees of freedom are now larger than before because we have taken 4 boys instead of only 2. And once again the differences between the 4 boys appear to be fully significant, but (with error assessed as above) the differences between the 4 tests are not significant.

Testing Reliability

Suppose that Tom, Dick, Harry and George are the names of four examiners marking test performances by four boys in the *same* subject. Thus the names of the rows down the left-hand margin of the table are names of candidates taking the tests. We can now use the analysis of variance to measure the reliability or self-consistency of the whole examination. We could vary this by making the headings of the columns four component tests instead of four different examiners. The reliability coefficient is given by

$$r_{tt} = \frac{\bar{P} - \bar{E}}{\bar{P}}$$

where $\bar{P}$ is the mean square for pupils or candidates and $\bar{E}$ is the mean square for error based on the residuals.[1]

Testing the Significance of Group Factors (Interaction)

Problem

The foregoing are the simplest and commonest types of case in which the analysis of variance can be applied. We now proceed to introduce a further complication.

In estimating the variance for error, we assumed that the deviations of the 8 performances from the combined means of boy and test were random deviations. A glance at the figures headed 'deviations for performances' on page 137 is sufficient to show that they are not random, but correlated. We must therefore treat them as containing yet another component — a bipolar component. This is technically termed *interaction*, because the type of boy tested 'interacts' with the type of test used, i.e. an

[1] This is developed by Burt in *The British Journal of Educational Psychology*, XV, pages 80-92. The use of factor analysis for a similar purpose is given in Burt, *Marks of Examiners*.

academic type of boy does well in the academic type of test, whether Arithmetic or English and, by comparison, badly in the practical type of test: conversely for the practical type of boy.

This *bipolar component* we can assess by averaging the deviations in each column, reversing the signs of the last two to prevent the totals adding up to zero. We then calculate the deviations about these further averages. Thus the variance of the deviations for performances can itself be analysed along the same lines as before.

(4) *Deviations for Bipolar Component*					(4) *Squares of Deviations*				
Arithmetic	+ 3	+ 4	- 1 5	- 5 5	9	16	2 25	30 25	57 5
English	+ 3	+ 4	- 1 5	- 5 5	9	16	2 25	30 25	57 5
Drawing	- 3	- 4	+ 1 5	+ 5.5	9	16	2 25	30 25	57 5
Handwork	- 3	- 4	+ 1 5	+ 5 5	9	16	2 25	30 25	57 5
Square Sum					36	64	9	121	230 0

(5) *Deviations for Error*									
Arithmetic	1	- 1	+ 0 5	- 0 5	1	1	0 25	0 25	2 5
English	1	- 1	- 0 5	+ 0 5	1	1	0.25	0 25	2 5
Drawing	0	0	+ 0.5	- 0 5	0	0	0 25	0.25	0 5
Handwork	0	0	- 0 5	+ 0 5	0	0	0 25	0 25	0 5
Square Sum					2	2	1	1	6

The degrees of freedom for the 'bipolar component' will evidently be 3; and those for the 'deviations for error' will evidently be 6. We have thus split what we previously assumed to be 'error' into two components. Note that both the square-sum and the degrees of freedom now obtained add up to those previously assigned to 'error' in the table of the analysis of variance for four boys.

We must now analyse the total variance afresh.

(In setting out tables like the following the beginner finds it best to set the obtained figure first, the degrees of freedom next, and the calculated or textbook figures last, since that is the order of working. The experienced worker, however, will put the degrees of freedom first, since they really indicate the structure and fundamental conditions of the analysis.)

ANALYSIS OF VARIANCE: (WITH FOUR COMPONENTS)

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*
Boys	736	$4 - 1 = 3$	$245.\dot{3}$
Tests	416	$4 - 1 = 3$	$138.\dot{6}$
Interaction	230	$4 - 1 = 3$	$76.\dot{6}$
Error	6	$16 - 10 = 6$	1.0
Total	1388	$16 - 1 = 15$	(92.5)

The observed and expected variance ratios may be tabulated as follows. The divisor is now 1.0 in every case.

VARIANCE RATIOS

Source	*Ratio*	*Observed*	*Degrees of Freedom*	*Expected* 5%	1%
Boys	F_b	$245.\dot{3}$	3 and 6	4.76	9.78
Tests	F_t	$138.\dot{6}$	3 and 6	4.76	9.78
Interaction	F_i	$76.\dot{6}$	3 and 6	4.76	9.78

Thus, when we allow for the fact that the tests are highly correlated, and thus confirm one another far more strongly than a random set of tests, the differences between boys, between tests, and between types of boy (or test) appear highly significant.

Application to Factor Analysis

It will now be seen that we have demonstrated the statistical significance of (i) the 'general factor' of average ability, and (ii) the 'group factor' of academic versus practical ability. Thus, provided the factor-measurements are obtained by simple averaging, we have found a convenient method for testing the significance of factors.

(The high significance thus obtained with a sample consisting of 4 boys only may seem surprising. But the correlations are equally high. Thus, the observed correlation for Arithmetic and English is .99 and the residual correlation .92. Now with 4 items the 1% level is .99 and the 5% level .90. But we have not one

correlation but 6 in each case, though not all 6 residual correlations will be independent. Thus the rough test of significance applied to the correlations confirms the more precise test obtained by analysing variance. However, it should be remembered that the figures given in this example are purely artificial, chosen to simplify the mental arithmetic, rather than to illustrate the kind of figures actually obtained.)

Interaction

When planning a research which will involve the analysis of variance the 'factors' are chosen not so much because they operate independently but because they can be controlled and measured. Thus it is necessary to devise methods of research wherein the joint effects of the varying factors may be compared with their isolated effects, and it is possible that the joint effect will not be the mere sum of the respective effects. We can adapt the methods given by Fisher in his *Design of Experiments* where the investigations concerned agriculture (manuring of fields, rotation of crops, etc.) to our educational problems. Much investigation remains to be done on suitable teaching methods for children of various ages and capacities and in various subjects. We might use (*a*) oral methods alone, (*b*) film strip, (*c*) cinema film, (*d*) practical work and exercises, and (*e*) a combination of two or more of these methods. We might expect that combinations of the methods might be more effective than the use of a single method.

In the analysis of variance what is known as error is the combined effect of various influences which either cannot be or are not controlled in the investigation. Certain precautions must be observed in order that we can estimate this error. With small sampling techniques it is necessary to secure the *replication* or repetition of individual items with similar factorial content. Where the 'interactions' are known or can be shown to be significant they may be used to measure error. Secondly, within the conditions imposed by the experimental design the items should be assigned at random. Randomisation may be secured by a mechanical method such as tossing coins, drawing cards or by using sets of random numbers. Fisher used the name 'randomised

blocks' for an experimental design which involved these principles. Eight blocks of land are selected and each is divided into five plots. Five varieties of a particular kind of crop, or five types of fertiliser are assigned at random to each plot. We could translate this into a research in education by testing the relative merits of five different methods of training. Such problems as the methods of teaching various processes in arithmetic, improving memorization or treating delinquents would be susceptible to such treatment. Obviously the children to be studied will differ according to home and school environment and thus the children used in the investigation are chosen from eight schools. Children of about the same age are picked at random from the schools and a different method of training is allotted at random to each individual. In analysing the results there will be only one criterion of classification — that according to training or treatment.

But if the number of performances is large enough the number of ways in which they are classified or cross-classified may be increased from two to three or more.

Example: We wish to investigate the efficacy of four different training methods (e.g. the remedial teaching of backward spellers). Four boys are selected and all four will be subjected to all the four methods. To obviate possible differences arising from the test words used in the experiment, all the words will have to be taught by all the methods. It is possible, even probable, that the order in which a boy is taught by the different methods may make some difference to the result. For instance, if he starts with a phonic method and goes on to a copying method, the latter might be helped by the former. Again, if he starts the week with a phonic method and goes on to the others on subsequent days this might affect the results. Thus, as far as can possibly be managed it is necessary so to arrange the order that, with one boy or another, each method follows and precedes every one of the others.

The following arrangement, which meets these requirements, is known as the *Latin Square* as the Roman or Latin capitals A, B, C and D represent the four methods. When a further classification is necessary Greek letters are used and the arrangement is then known as a *Graeco-Latin Square*.

ARRANGEMENT OF TEACHING METHODS IN A LATIN SQUARE

Order	*Tom*	*Dick*	*Harry*	*George*
1	A	B	C	D
2	B	D	A	C
3	C	A	D	B
4	D	C	B	A

We will now express the marks in the tests designed to examine the teaching methods. For convenience in analysis these have been arranged in the form of deviations from the general mean.

RESULTS OF TEACHING

Test Material	*Tom*	*Dick*	*Harry*	*George*	*Total*	*Average*	*Square*
I	26	15	− 3	− 10	28	7	49
II	22	5	− 1	− 18	8	2	4
III	− 10	1	− 9	2	− 16	− 4	16
IV	− 2	− 9	− 7	− 2	− 20	− 5	25
Total	36	12	− 20	− 28	0	0	94
Average	9	3	− 5	− 7	0		
Square	81	9	25	49	164		

To calculate the averages for each training method we rearrange the figures in each column as follows:

Teaching Method	*Tom*	*Dick*	*Harry*	*George*	*Total*	*Average*	*Square*
A	26	1	− 1	− 2	24	6	36
B	22	15	− 7	2	32	8	64
C	− 10	− 9	− 3	− 18	− 40	− 10	100
D	− 2	5	− 9	− 10	− 16	− 4	16
Total	36	12	− 20	− 28	0	0	216

From each figure in the last table but one we now subtract the sum of the appropriate averages for (i) the boy, (ii) the test material, and (iii) the teaching method. We obtain the following residuals:

RESIDUALS AND THEIR SQUARES

Test Material	*Tom*	*Dick*	*Harry*	*George*	*Total*	*Tom*	*Dick*	*Harry*	*George*	*Total*
I	4	− 3	5	− 6	0	16	9	25	36	86
II	3	4	− 4	− 3	0	9	16	16	9	50
III	− 5	− 4	4	5	0	25	16	16	25	82
IV	− 2	3	− 5	4	0	4	9	25	16	54
Total	0	0	0	0	0	54	50	82	86	272

The sums of the squares are tabulated below. In entering those for each of the means we have multiplied the squares from a single column or row by the number of columns or rows (in this case 4), since the means are repeated in each column and in each row.

ANALYSIS OF VARIANCE (LATIN SQUARE)

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*
Boys	3	656	$218.\dot{6}$
Test Material	3	376	$125.\dot{3}$
Teaching Methods	3	864	288.0
Residuals	6	272	$45.\dot{3}$
Total	15	2168	

VARIANCE RATIOS, OBSERVED AND EXPECTED

Source	*Observed*	*Degrees of Freedom*	*Expected* 5%	1%
Boys	$\frac{218.\dot{6}}{45.\dot{3}} = 4.82$	3 and 6	4.76	9.78
Test Material	$\frac{125.\dot{3}}{45.\dot{3}} = 2.76$	3 and 6	4.76	9.78
Teaching Methods	$\frac{288.0}{45.\dot{3}} = 6.35$	3 and 6	4.76	9.78

The differences in the effects of teaching are fully significant but those for the boys are only just over the borderline. There is no discernible difference in the different types of teaching material.

With a more elaborate experiment we could study the interactions, that is, the differences in effect of teaching methods on particular types of pupil or test material. It has been assumed in the above example that the 'interaction' can be taken as a measure of error for the main effects.

Methods of Working

In actual practice it will involve considerable labour to work with the actual deviations, for the means will usually involve decimal fractions. The following procedure will make the arithmetical work simple and mechanical. It will be illustrated from the problem on page 144 involving three criteria. Here are the steps of the process:

1. Find the totals of the rows and the columns, and the grand total.

2. Divide the totals by the number in the corresponding row, column or table.

3. Multiply each total by the corresponding mean.
This may be done by a calculating machine, but if one is not available, square the means, and multiply by the number of items on which each mean is based. (The result is obviously the same, but the 'total × mean' method avoids any mistakes in multiplying the squares, when the number of rows differs from the number of columns.)

4. Add the products.

5. With the Latin Square rearrange the rows and find the 'totals × means' as before.

6. Square each figure in the first table and find the grand total of the squares.

7. From each of the four totals thus obtained, subtract the product of the grand total by the grand mean. The results are the square-sums for the various means and the total square-sum.

8. To find the square-sum for the residuals, subtract the sum of the three square-sums for the means from the total square-sum. The final result can be checked by directly calculating the squares for the residuals, at least approximately.

WORKING METHOD. STEPS I, II, III AND IV

Test Material	*Tom*	*Dick*	*Harry*	*George*	*Total*	*Mean*	*Product*
I	A 46	B 35	C 17	D 10	108	27	2916
II	B 42	D 25	A 19	C 2	88	22	1936
III	C 10	A 21	D 11	B 22	64	16	1024
IV	D 18	C 11	B 13	A 18	60	15	900
Total	116	92	60	52	320	80	6776
Mean	29	23	15	13	80	20	
Product	3364	2116	900	676	7056		6400

STEP V

A	46	21	19	18	104	26	2704
B	42	35	13	22	112	28	3136
C	10	11	17	2	40	10	400
D	18	25	11	10	64	16	1024
Total							7264

STEP VI

I	2116	441	361	324	3242
II	1764	1225	169	484	3642
III	100	121	289	4	514
IV	324	625	121	100	1170
Total	4304	2412	940	912	8568

STEP VII

	Crude Square Sum		*Correction Term*		
Boys	7056	—	6400	=	656
Test-Material	6776	—	6400	=	376
Teaching Methods	7264	—	6400	=	864
Total	8568	—	6400	=	2168

STEP VIII

Square Sum for Residuals $2168 - (656 + 376 + 864) = 272$

Such comparatively simple analysis may lead to more elaborate experimental designs such as those in which there may be two or three criteria of classification, one or two essential interactions and several items instead of only one in each sub-class.[1] The technique

[1] See Sir Cyril Burt's report on 'Teaching Backward Readers', *British Journal of Educational Psychology*, XVI.

may also be extended to the testing of simple and multiple regressions and their linearity. This is given in Mather, Chapters VIII and IX. It may also be applied to intra-class correlation (see Fisher, *Statistical Methods*) and to the analysis of covariance. The latter is necessary where the criteria of classification may be not independent but correlated. Suppose it is necessary to test alleged differences in educational attainments between children in various parts or towns of a county at a transfer examination. It may be that the age composition may vary from one part to another. Regression must then be used to eliminate the effects of differing age. This is best done by analysing the covariance as well as the variance. The method is given in Snedecor, Chapter VIII.

The works of Fisher, Snedecor, Yule and Tippett mentioned in the bibliography may be consulted for more advanced work on the analysis of variance.

APPENDIX I

GRAPHS AND GRAPHICAL METHODS. THE DIFFERENTIAL CALCULUS AND TRIGONOMETRICAL FUNCTIONS

GRAPHICAL methods of expression will prove very helpful in simple statistical investigations. In fact, for those who have only the slightest knowledge of mathematics they will often prove to be the only means of dealing with the results of an investigation, lists of scores and so on. Even where the investigator is well equipped mathematically graphical method still remains as the best means of recording and interpreting results, in many cases.

Graphs make an immediate appeal to the eye. Even where there is little 'aptitude for figures' the visual image is the one above all others, which can be most easily remembered, analysed and interpreted.

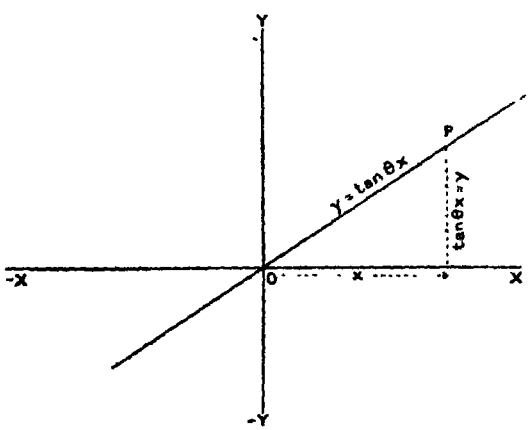
Graphs give a picture of the variation of one quantity with another, and properly interpreted the graph will provide a clue to the extent and nature of this variation.

Unless the investigator knows something of the calculus, of exponentials, etc., the graph is often the only means of representing the variation. Finding the areas enclosed by graphs is an easy way of 'integrating'; tangents drawn to points on curved graphs anticipate the process of 'differentiating'. Maximum and minimum points are easily seen and interpreted. With a graph, *interpolation* is possible, that is, intermediate values between the plotted points may be found. A curve or line may be extended by having regard to its general shape and hence finding further values which are outside the range of the points that are plotted. This is known as *extrapolation*. The processes of interpolation and extrapolation are not to be undertaken lightly. In the former case intermediate values should be found by experiment and observation particularly where a curve turns sharply. In the latter case the continuation of a line is a very risky procedure for

factors may come into play which alter the general trend and in psychological investigations these 'tails' may have considerable significance. Interpolation and extrapolation should be applied on the merits of each case and then with care and reticence.

A point *xy* may be fixed on a plane surface by referring it to two axes. It is convenient to draw these as straight lines at right angles. If the horizontal and vertical axes divide the graph paper into four equal parts we can provide for an equal number of x and negative x values and of y and negative y values. If we are only concerned with positive values of x and y it will suffice to draw the axes respectively at the bottom and at the left side of the paper. Distances are measured from the origin which is the point *o* where the axes intersect, and it is *conventional* to regard values measured to the right and upwards as positive and those to the left and downwards as negative. To plot a point *xy* it is necessary to measure along the x axis a distance x and upwards a distance y. It is necessary to consider carefully what scales can be employed for both x and y values, in other words, how many units of x and y are represented by a division on the graph paper.

If a straight line is drawn on the graph paper it will contain a series of points which represent values of x and y which are related together in a simple way. x and y are connected together



in terms of a simple equation, appropriately called a *linear equation.* The value of y is dependent on that of x: y is known as the *dependent variable*, and x the *independent variable.* y becomes a function of x and is sometimes written $y = f(x)$.

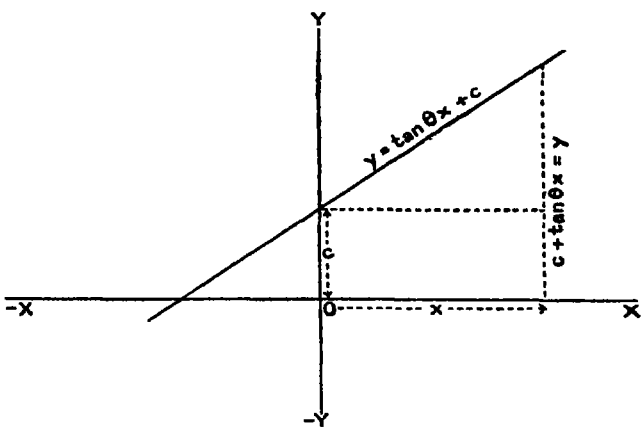
Let us first consider a straight line drawn through the origin o and at an angle θ (theta) with the axis of $x(ox)$.

Consider *any* point P on the line.

Its *co-ordinates*, that is its x and y values, are related together by

$$\frac{y}{x} = \tan\theta \qquad \text{or } y = x\tan\theta$$

The slope of the line can thus be thought of as the tangent of the angle which the line makes with the axis of x.[1] The equation of this line has already been given: it is $y = x\tan\theta$ and this connects all the x and y values on the line.



When the line does not go through the origin but meets the axis of y at the point l cutting off a piece oc (c) on it, it will readily be seen that the equation of the straight line is $y = x\tan\theta + c$ for every y value corresponding to an x in the previous equation of the line through the origin will have to be increased by the intercept c on the axis of y.

[1] See page 157.

Any equation which can be put in the form $lx + my + n =$ where l, m and n are independent of x and y can be represented o a graph as a straight line.

In this case, the slope of the line $= -\frac{l}{m}$

and the intercept on the axis of $y = -\frac{n}{m}$

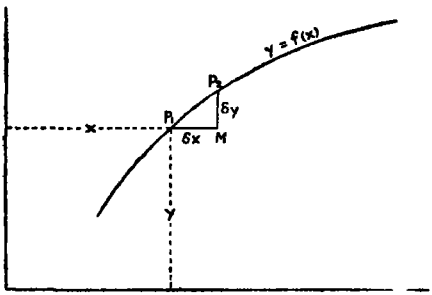
A linear relationship is said to exist between two sets of measures if a straight-line graph is yielded when points representing corresponding sets of values are plotted and joined.

The use of straight-line or other graphs as ready reckoners, conversion tables, etc , needs no stressing.

A few words should be said about regression lines. The line $y = rx$ gives the regression of y on x and $x = ry$ gives the regression of x on y. Where $r = 1$ (perfect correlation) the line $y = x$ goes through the origin and makes an angle of 45° with both axes. The older school of statisticians would say that when correlation was perfect there was no regression, but some writers make r the correlation coefficient (and slope of the regression line) a direct measure of the regression. From the context it is usually easy to see what a writer intends to convey. Regression gives us a measure of the reliability of predicting the value of a measure by reference to that of another with which it is correlated to a greater or lesser degree.

The calculus is best approached by considering the graphs of curves. We may look upon differentiation as a process of measuring rate of change, curvature, etc., and integration as one of summation, the determination of areas, etc. Differentiation and integration may be regarded as one the reverse of the other. As these processes involve conceptions relating to infinity and infinitesimals care must be taken to see that these ideas are not given the form of absolute numbers.

Suppose the curved line represents a function $f(x)$ of x. Its equation is $y = f(x)$. Consider a point on the line P_1 whose co-ordinates are x and y. Further, take another point P_2 near to it with co-ordinates slightly larger $x + \delta x$ and $y + \delta x$, where δx



and δy (delta x and delta y) are small increments in the value of x and y respectively.

Now consider the small triangle P_2MP_1 with vertical side δy and base δx. Its hypotenuse P_1P_2 will approximate to a portion of the curve as δy and δx become smaller.

P_1P_2 will be a tiny part of a tangent to the curve as P_1 and P_2 approach one another.

The slope of this tangent $= \dfrac{\delta y}{\delta x}$

$$\begin{aligned} y &= f(x) \\ \therefore y + \delta y &= f(x + \delta x) \\ \therefore \delta y &= f(x + \delta x) - f(x) \\ \frac{\delta y}{\delta x} &= \frac{f(x + \delta x) - f(x)}{\delta x} \end{aligned}$$

It is necessary to utter a word of warning that the rigorous treatment of the calculus must be regarded as being beyond the scope of this short statement. $\dfrac{\delta y}{\delta x}$ is a true quotient obtained by dividing small but finite quantities δy and δx but when we proceed to the limit and obtain the differential coefficient $\dfrac{dy}{dx}$ this must not be regarded as a fraction but as an operator $\dfrac{d}{dx}$ acting on y. The differential coefficient of a function is spoken of as its first derivative and is represented by $f_1(x)$.

A simple example will show this method in use

$$\text{Suppose } y = x^2$$
$$y + \delta y = (x + \delta x)^2$$
$$y + \delta y = x^2 + 2x\delta x + \delta x^2$$
$$\therefore \delta y = x^2 + 2x\delta x + \delta x^2 - x^2$$
$$= 2x\,\delta x + \delta x^2$$
$$\therefore \frac{\delta y}{\delta x} = 2x + \delta x$$

Now making δy and δx smaller and smaller

$$\frac{dy}{dx} = 2x$$

(This is where this method, though simple, lacks rigour, for we assume that δx vanished but that $\frac{\delta y}{\delta x}$ becomes $\frac{dy}{dx}$. The above method might be regarded as a useful demonstration rather than a proof.)

To find the differential coefficient or the derivative for x^n we need to keep in mind the binomial expansion for $(x + a)^n$

$$(x + a)^n = x^n + nx^{n-1}a + \frac{n(n-1)}{1 \times 2} x^{n-2} a^2$$
$$+ \frac{n(n-1)(n-2)}{1 \times 2 \times 3} x^{n-3} a^3 \ldots\ldots a^n$$
$$y = x^n$$
$$y + \delta y = (x + \delta x)^n$$
$$= x^n + nx^{n-1}\,\delta x + \frac{n(n-1)}{1 \times 2} x^{n-2}\,\delta x^2 + \ldots$$
$$\therefore \delta y = nx^{n-1}\,\delta x + \frac{n(n-1)}{1 \times 2} x^{n-2}\,\delta x^2 + \ldots$$
$$\frac{\delta y}{\delta x} = nx^{n-1} + \frac{n(n-1)}{1 \times 2} x^{n-2}\,\delta x + \ldots$$

term containing higher power of δx.

$$(x + a)^n = x^n + nx^{n-1}a + \frac{n(n-1)}{1 \times 2}x^{n-2}a^2 + \frac{n(n-1)(n-2)}{1 \times 2 \times 3}x^{n-3}a^3 \ldots . a^n$$

To find $\frac{dx^n}{dx}$

$y = x^n$

$y + \delta y = x^n + nx^{n-1}\delta x + \frac{n(n-1)}{1 \times 2}x^{n-2}\delta x^2 + \ldots .$

$\delta y = nx^{n-1}\delta x + \frac{n(n-1)}{1 \times 2}x^{n-2}\delta x^2 + \ldots .$

$\frac{\delta y}{\delta x} = nx^{n-1} + \frac{n(n-1)}{1 \times 2}x^{n-2}\delta x +$ terms containing higher powers of δx.

Proceeding to the limit

$\frac{dy}{dx} = nx^{n-1}$ as terms containing δx and its powers vanish in the limit.

As the differential coefficient gives a measure of the slope of the curve it will be equal to 0 where the curve has no slope, that is to say at the points of the curve where the tangents are horizontal.

Thus, we find values of x which correspond to maximum or minimum values of the function by equating the differential coefficient to zero and solving the equation.

This method will not distinguish between maximum and minimum values but it can readily be seen that, as we trace out a curve, a tangent to the moving point will turn in a clockwise direction as we approach and pass a maximum value and it will turn in an anticlockwise direction as we approach and pass a minimum.

Thus, a further process of differentiation (double differentiation) will give us a clue to the recognition of maxima and minima.

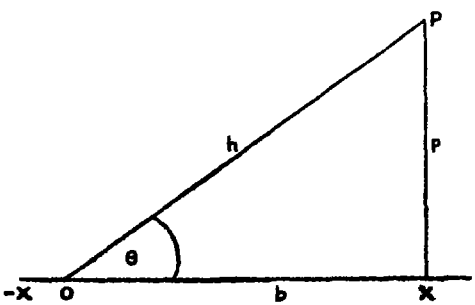
If the second differential coefficient $\frac{d^2y}{dx^2}$ has a positive value the point concerned will be a minimum and if it has a negative value the point will be a maximum.

Differentiation	Integration
$\frac{d}{dx} x^{n+1} = (n+1)\,x^n$	$\int x^n\,dx = \frac{x^{n+1}}{n+1}$
$\frac{d}{dx} \log_e x = \frac{1}{x}$	$\int \frac{dx}{x} = \log_e x$
$\frac{d}{dx} \cos x = -\sin x$	$\int \sin x\,dx = -\cos x$
$\frac{d}{dx} \sin x = \cos x$	$\int \cos x\,dx = \sin x$
$\frac{d}{dx} \tan x = \sec^2 x$	$\int \sec^2 x\,dx = \tan x$
$\frac{d}{dx} \cot x = -\operatorname{cosec}^2 x$	$\int \operatorname{cosec}^2 x\,dx = -\cot x$
$\frac{d}{dx} e^x = e^x$	$\int e^x\,dx = e^x$

$\frac{d}{dx}$ should be regarded as an operator and *not* as a fraction.

A constant which can be determined from the practical nature of the problem and the given data has to be added in each case. This is obvious when it is remembered that integration is the reverse of differentiation and that the differential coefficient of a constant is zero.

Trigonometrical Functions of an Angle



Consider the right-angled triangle POX with angle POX $= \theta°$. PX (perpendicular) $= p$, OX (base) $= b$, OP (hypotenuse) $= h$.

sine θ (sin)	$= \frac{p}{h}$	cotangent θ (cot)	$= \frac{b}{p}$
cosine θ (cos)	$= \frac{b}{h}$	secant θ (sec)	$= \frac{h}{b}$
tangent θ (tan)	$= \frac{p}{b}$	cosecant θ (cosec)	$= \frac{h}{p}$

It will readily be seen using the properties of a right-angled triangle that each of these functions may be calculated by knowing any one of the others. The following relationships are most important

$$\tan\theta = \frac{\sin\theta}{\cos\theta}, \quad \sin^2\theta + \cos^2\theta = 1.$$

$$\cot\theta = \frac{1}{\tan\theta}, \quad \sec\theta = \frac{1}{\cos\theta}, \quad \operatorname{cosec}\theta = \frac{1}{\sin\theta}.$$

$$\cos(90° - \theta) = \sin\theta \quad \sin(90° - \theta) = \cos\theta.$$

The angle θ must not be regarded as an angle limited to less than a right angle. A triangle of reference POX may be drawn by dropping a perpendicular PX from a point P on the line OP generating the angle on to the axis of X, $-$ X.OX. Although the tables only give angles between 0° and 90° the trigonometrical functions for other angles may be calculated by arranging them as $(180° - \theta)$, $(180° + \theta)$, $(360° - \theta)$ where θ is an angle less than 90° which can be found from the tables. The following diagram shows when it is necessary to change the sign of the function found in the tables. Angles are measured in an anticlockwise direction and the complete round of angles (360°) is divided into four quadrants

$(180° -)$ sine + cosec +	All +
$(180° +)$ tan + cot +	$(360° -)$ cosine + sec +

or in the mnemonic form by using the word CAST: $\begin{array}{c|c} S & A \\ \hline T & C \end{array}$

It may be useful to remember that:

$\sin 0° = 0$	$\cos 0° = 1$	$\tan 0° = 0$
$\sin 30° = \frac{1}{2}$	$\cos 30° = \frac{\sqrt{3}}{2}$	$\tan 30° = \frac{\sqrt{3}}{3}$
$\sin 45° = \frac{1}{\sqrt{2}}$	$\cos 45° = \frac{1}{\sqrt{2}}$	$\tan 45° = 1$
$\sin 60° = \frac{\sqrt{3}}{2}$	$\cos 60° = \frac{1}{2}$	$\tan 60° = \sqrt{3}$
$\sin 90° = 1$	$\cos 90° = 0$	$\tan 90° = \infty$ (infinity)

Sometimes mental tests, mental 'factors', etc., are represented as vectors, that is, straight lines at an angle to one another. The correlation coefficient between the quantities represented by any two lines is given by the cosine of the angle between them. The projection of one line upon another is equal to the length of the first line multiplied by the cosine of the angle between the lines. (Do not confuse this with regression and remember that the 'slope' of a line is given by the *tangent* of the angle which it makes with an *axis of reference*.)

Factors, etc., represented by vectors at right angles are obviously uncorrelated ($\cos 90° = 0$) and they are said to be *orthogonal*.

Factors, etc., represented by vectors which are not at right angles contain some measure of correspondence (the cosine of the angle between them is not zero). These are said to be *oblique* factors.

This useful idea can be extended from two dimensions to three (and analytically without trying to conceive models to 4 or more. The geometry of hyperspace can be used for dealing with more than 3 factors which are represented by vectors). Three orthogonal factors can be thought of as lying along the edges of a rectangular box and meeting at one of its corners. A number of oblique factors could be drawn as lines in space radiating from a point. If an arbitrary line were taken to represent the first factor

the other lines could be imagined to fit into their relative positions by taking the correlation coefficient between each pair, finding the angle of which it is the cosine and fitting in the line accordingly. With three lines this involves a simple principle of solid geometry but with four or more analytical methods using algebra and trigonometry may have to suffice. Angles are not always given in degrees, and it is often more convenient to think of them in radian measure.

$$\begin{aligned} 2\pi \text{ radians} &= 360^\circ \\ \pi \text{ radians} &= 180^\circ \\ 1 \text{ radian} &= \frac{180^\circ}{\pi} \end{aligned}$$

When the symbol π appears in formulae used in psychological and educational statistics it usually refers to an angle of two right angles or 180°.

APPENDIX II

THE USE OF THE SLIDE-RULE[1]

THE slide-rule, which dates from about the same period as that of the invention of logarithms, is really a simple instrument working on logarithmic principles. To multiply two numbers we add their logarithms. If, therefore, we have two scales whose distances and divisions are measured out in the lengths of the logarithms which they represent it is easy to see that numbers may be multiplied by adding these logarithmic lengths by means of two scales one of which is capable of sliding against another. Division may be performed by subtracting these logarithmic lengths, squaring by doubling and finding a square root by halving and so on. In our work the slide-rule is particularly useful when each of a set of numbers has to be multiplied (or divided) by a factor, as for instance in reducing a set of marks from one maximum to another. One setting of the rule is all that is required and the reduced marks may be read off directly from the rule.

Although most work in educational and psychological statistics does not call for the full resources of the instrument such as is used by engineers, it is worth while to acquire a good one, which will cost from 30s. to £3. The beginner need not feel overwhelmed by the amount of metrical material compressed into one scale. If any difficulty arises it will suffice to make a simple slide-rule by gumming two strips of logarithmic graph paper to two ruler-like pieces of wood respectively which can be made to slide against one another and may be kept together by a couple of small elastic bands. No difficulty is expected, however.

Finding Numbers

The front face of the ordinary 10-inch slide-rule consists of two pairs of scales; the upper ones usually are called the A and B scales and the lower pair are known as the C and D scales.

[1] See also the section on Logarithms in *The Teaching of Arithmetic and Elementary Mathematics*, by the author.

Any number of whatever reasonable magnitude can be located on the slide-rule, because the first mark can be called 1, 10 or 100 as required. The sub-division of units sometimes gives difficulty at first but since there are only three different variations to learn these should be mastered at the outset.

If we call the first mark on the A scale 10, the number 11 is to be found five graduations (division marks) further along, the space between 10 and 11 is divided into five parts, with graduation marks at 10·2, 10·3, 10·6, 10·8 leaving any smaller divisions to be estimated as required. This method of marking continues until 20 is reached, after which the spaces between the whole numbers are not large enough to allow five divisions, so from thence onwards the units are only cut in half. From 50 to 100 there is not even room for this to be done and the units are no longer subdivided.

On the D scale there is more room as 'smaller' numbers are involved. If the beginning is called 10, the number 11 is found ten marks further along, the intermediate values being 10·1, 10·2, etc., to 10·9 and this system is continued up to 20 From 20 to 40 the units have five divisions each, e.g., 20·2, 20·4, 20·6, 20·8, after which there is only sufficient room for half divisions to be shown.

If a 10-inch slide-rule is examined carefully so that these facts are appreciated facility in finding and reading numbers will soon follow.

It is always worth while to perform rough mental calculations of the answer as this will help to find the correct place for the decimal point.

1. *Multiplication*

Example: 14·6 × 3·2 (approximate value 50). Put B. 1 (the beginning of the B scale) against one of the numbers on the A scale. Locate the second number on the B scale and read off the product from the A scale immediately above the B scale number. The fine vertical line of the transparent window of the sliding cursor may help in reading a number on one scale which is exactly in line with a number on the other.

In effect, in this process of multiplication a piece of the A scale has been added to a piece of the B scale and, as the numbers are multiplied together by adding their logarithmic lengths, the total length indicates the products of the two numbers.

2. *Division*

Example: 43·6 ÷ 19·8 (estimated approximate value 2).

Place the divisor 19·8 on B scale immediately under the dividend 43·6 on the A scale The quotient may be read off on the A scale immediately above B 1. In division a piece of Scale B is subtracted from a piece of Scale A To divide two numbers we subtract their logarithms.

Both multiplication and division can be performed on the C and D scales. The results can usually be estimated to a greater degree of accuracy owing to the larger divisions, but working is generally a little slower than with the A and B scales.

3. *Conversion and Reduction*

These processes are equivalent to multiplying or dividing the given number by a certain factor. It will be seen that division by a number is equivalent to multiplication by the reciprocal of the number, e.g. division by 12 is equivalent to multiplication by $\frac{1}{12}$ or ·0833. Each case must be considered on its merits, that is, whether it is easier to multiply by a factor or divide by its reciprocal. *Example*: To convert marks given with a maximum score of 80 to a maximum of 100. This is equivalent to multiplying each mark by $\frac{100}{80}$ or 1·25. For ease of working it is better to put B1 opposite to 1·25 on the A scale and read off the result on the A scale immediately above the given number on the B scale. After the initial setting no further movement of the scale will be required for the whole set of marks.

The conversion of marks from a maximum of 100 to one of 80 need not be regarded as a division but rather as a multiplication by the factor ·8.

Squaring Numbers

Find the number on the D scale. Its square lies immediately above it on the A scale. Use the cursor. The scales all remain at 'zero' position.

Finding Square Roots

Find the number on the A scale. Its square root lies immediately below it on the D scale. Now any number which is given in figures without a decimal point will appear to have a choice of one of two square roots (quite apart from negative roots), e.g. the square root of 4·0 is 2·0 but that of 40 is 6·3 Thus there are two positions for any number on the A scale, and the correct one must be chosen with reference to the size of the given number according to the following rule. For numbers with an *odd* characteristic use the right-hand part of the A scale. For numbers with an *even* characteristic use the left-hand part. The characteristic is one less than the number of digits to the left of the decimal point, and if negative is one more than the number of noughts *immediately* to the right of the decimal point, e g.

3167	characteristic	3	odd
316·7	characteristic	2	even
9·6	characteristic	0	even
·3076	characteristic	— 1	odd
·0003001	characteristic	— 4	even

In using tables of square roots the same principle applies, but it is usually sufficient to make a rough mental estimate of the required value and this will determine which of the two given numbers is required.

APPENDIX III

PASCAL'S TRIANGLE AND THE NORMAL CURVE OF DISTRIBUTION

SUPPOSE that we toss a penny a large number of times. In the long run heads and tails will be about equally divided and the distribution will be in the proportion

H	T
1	1

If we toss two pennies there will be three possibilities: two heads, two tails, one head, one tail, in the proportion:[1]

HH	HT HT	TT
1	2	1

With three pennies there will be four possibilities: three heads, three tails, one head two tails, one tail two heads in the proportion

HHH	HHT	HTT	TTT
1	3	3	1

and so on. Although we do not find these proportions strictly observed unless we take inconveniently or impossibly large numbers of cases these figures represent the probabilities of the distributions of each particular showing of heads and tails.

This at once suggests to us that it may be useful to consider the numbers arising when we continue to multiply 11 by itself, that is, the powers of 11

$(11)^1$	11
$(11)^2$	121
$(11)^3$	1331
$(11)^4$	14641

[1] Students of biology will note that these are the proportions of offspring showing distinct transmissible characteristics in the simplest application of Mendel's laws, e.g. in the second generation of peas in the crossing of long and short peas, pure long peas, impure long peas and pure short peas were in the proportion 1 2. 1 respectively.

In building up Pascal's triangle we must continue the powers of 11 without carrying additions above 10 into a higher column. Those who are familiar with the binomial theorem will see that the above continuous multiplication by 11 gives the coefficients in the binomial expansion $(1 + x)^n$ of the ascending powers of x.

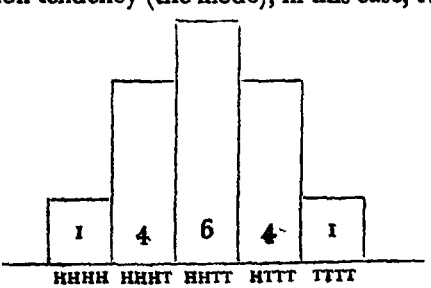
Thus $(1 + x)^4 = 1 + 4x + 6x^2 + 4x^3 + x^4$ by the expansion of $(11)^4$ or 1 4 6 4 1.

If it can be imagined that we continue Pascal's triangle to the limit making the number of the power n sufficiently large we should arrive at the exponential curve known as the probability curve or the curve of error. If instead of thinking of the smooth curve which is reached in the limit, let us imagine the histogram given at the bottom of this page.

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1

Pascal's triangle

It will readily be seen that the area of the whole figure represents the total number of cases i.e. $1 + 4 + 6 + 4 + 1 = 16$, the height of any column the frequency for each distribution of heads and tails and the distance from the centre point of the horizontal line (the x distance) the degree of departure from the central or most common tendency (the mode), in this case, two heads and



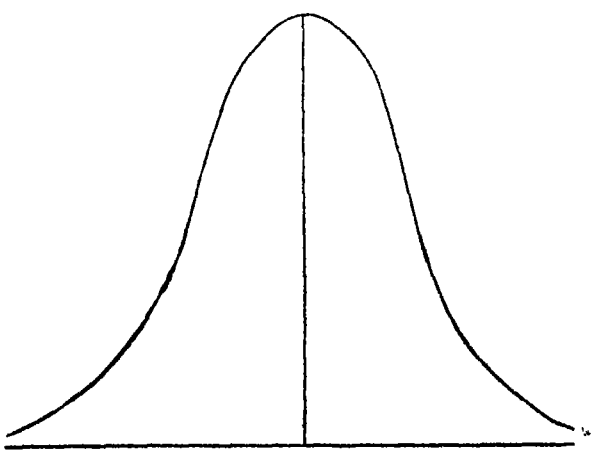
two tails. The chance that a single throw of four coins will give a particular number of heads and tails is given by the area of the column concerned compared with that of the whole figure., e.g. the chance of throwing four heads (or four tails) is 1 in 16.

If we now return to consider the histogram 'smoothed out' and its area representing a large number of cases, it is easy to appreciate that the probability that a measure (x) will lie at a certain distance from the central point is given by the ratios of the area of the tail of the curve beyond that point and the area of the remainder of the curve cut off by an ordinate through the point. In some cases (and these should be obvious) it will only be necessary to consider *one* half of the curve, that is, one or other of the halves on either side of the central line.

Some Properties of the Normal Curve of Distribution

This curve is also spoken of as the curve of error, the Gaussian curve or the curve of probability for reasons which we have already mentioned.

The curve is a member of the family of exponential curves, that is, it is related to the *growth function e*. The exponential function e^x has a rate of growth equal to itself i.e. $\frac{d\,e^x}{dx} = e^x$.



The curve may be defined as a frequency curve whose height at any point is inversely proportional to the antilogarithm of half the square of the distance, measured in terms of the standard deviation as the unit, of that point from the mean.

The formula for the curve is $y = y_0 e^{\frac{-x^2}{2\sigma^2}}$ where x and y are points on the curve with respect to o the central point on the x axis and y_0 is the 'height' of the curve at its central point, that is, the distance which it cuts off along the y axis.

σ is the standard deviation.

If this is large the curve is flat at the top and is said to be *platykurtic*. If this is small the curve is sharp and pointed and is said to be *leptokurtic*.

The degree of curvature is spoken of as *kurtosis*.

For our purpose we must regard y as a frequency of a score x which is referred to the average as zero.

We will differentiate the function representing the curve of normal distribution, written as,

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

where N is the number of cases in the distribution and σ is the standard deviation.

Let us write $\frac{N}{\sigma\sqrt{2\pi}} = c$ a constant

$$y = c\, e^{-\frac{x^2}{2\sigma^2}}$$

$$\frac{dy}{dx} = c\, e^{\frac{-x^2}{2\sigma^2}} \cdot \frac{d\left(-x^2 2\sigma^2\right)}{dx} = \left(c\, e^{\frac{-x^2}{2\sigma^2}}\right)\left(-\frac{2x}{2\sigma^2}\right)$$

$$= -\frac{cx}{\sigma^2} e^{\frac{-x^2}{2\sigma^2}} = \frac{-cx}{\sigma^2 e^{\frac{x^2}{2\sigma^2}}}$$

If we substitute $x = 0$ in this derived function (first differential coefficient) it vanishes.

Thus, this represents a value where the curve is at a maximum or minimum value. It is easy to see that this is actually a *maximum.*

Let us try to find other points where the curve has a maximum or minimum value, i.e. where it is horizontal.

Equate the first derived function to zero.

$$\frac{-cx}{\sigma^2 e^{\frac{x^2}{2\sigma^2}}} = 0$$

Divide by $\frac{-cx}{\sigma^2}$ $\qquad \frac{1}{e^{\frac{x^2}{2\sigma^2}}} = 0$

$$\therefore e^{\frac{x^2}{2\sigma^2}} = \infty$$

Taking logs. $\qquad \log_e \left(e^{\frac{x^2}{2\sigma^2}}\right) = \infty$

$$\therefore \frac{x^2}{2\sigma^2} (\log_e e) = \infty$$

Now $\log_e e = 1$ $\qquad \frac{x^2}{2\sigma^2} = \infty$

$$\therefore x^2 = \infty$$

$$x = \pm \infty$$

Thus, the curve is horizontal at infinite distances from the central line. (It is necessary to give a word of warning about the above demonstration. We have used 'infinity' as though it were a number and this may lead to absurdities. The above is not a rigorous demonstration and it is wise to warn the student against using 'plus and minus infinity'. Here we have unfortunately had to sacrifice rigour for the sake of a simple demonstration.)

Students who have proceeded a little further with the calculus than we have done here will be able to continue and find the second derivative or differential coefficient of the function of the curve of normal distribution.

$$\frac{dy}{dx} = \frac{-cx}{\sigma^2} e^{\frac{-x^2}{2\sigma^2}}$$

$$\frac{d^2y}{dx^2} = \left(\frac{-c}{\sigma^2}\right)\left(e^{\frac{-x^2}{2\sigma^2}}\right) + \left(\frac{-cx}{\sigma^2}\right)\left(e^{\frac{-x^2}{2\sigma^2}}\right)\left(\frac{-x}{\sigma^2}\right)$$

$$= \frac{c(x^2 - \sigma^2)}{\sigma^4} e^{\frac{-x^2}{2\sigma^2}}$$

$$= \frac{N}{\sqrt{2\pi}} \frac{(x^2 - \sigma^2)}{\sigma^5} e^{\frac{-x^2}{2\sigma^2}}$$

It will be observed that at symmetrical points of the curve there are points of inflexion, that is, the convex curvature of the top part of the curve gives way to the concave lower portions on each side. The rate of curvature will obviously be zero at these points. We can find them by equating the second derivative of the function to zero:

$$\frac{c(x^2 - \sigma^2)}{\sigma^4} e^{\frac{-x^2}{2\sigma^2}} = 0$$

Dividing through by $\frac{c}{\sigma^4} e^{\frac{-x^2}{2\sigma^2}}$ $\qquad (x^2 - \sigma^2) = 0$

$$\therefore x^2 = \sigma^2 \qquad x = \pm \sigma$$

Thus the points of inflexion are at a distance σ from the central point.

Let us consider the curve drawn on such a scale that its area is unity. The total number of cases N given by the area of the curve will be represented by unit area.

At the centre point or origin where $x = 0$ the equation of the curve becomes

$$y = \frac{1}{\sqrt{2\pi}\,\sigma} e^{0} = \frac{1}{\sqrt{2\pi}\,\sigma}$$

Thus $\frac{1}{\sqrt{2\pi}\,\sigma}$ is the height of the curve at its maximum (its modal ordinate) or the intercept cut off by the curve on the axis of y.

The area of the curve $\frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{x^2}{2\sigma^2}}$ can be found by integration. The curve must be thought of as extending from an infinite distance to the left of the centre point to an infinite distance to the right.

The total area is given by

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-x^2}{2\sigma^2}}$$

which is equal to 1.

[If this exponential curve could be considered as a development from the expansion of the binomial $(\frac{1}{2} + \frac{1}{2})^n$ the sum of all the ordinates is 1 for $(\frac{1}{2} + \frac{1}{2})^n = 1^n = 1$.]

The expression representing the normal curve may be written

$$y = \frac{1}{\sigma} z$$

$$\text{where } z = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

From statistical tables we may find values of z* for various values of $\frac{x}{\sigma_x}$

If the curve has unit area and unit standard deviation $y = z$

$$\text{and } y = z = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

If N is the area of the curve the equation of the curve of normal distribution is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

It is often necessary to find the area of a curve which lies between the central line and a vertical line at a distance from the origin, or the area of the 'tail' of the curve beyond a given value of x. Tables are provided of the values of such areas in Chapter V. These are usually denoted by q. It will be seen that the sum of these two areas is equal to the total area of the curve on one *or* other side of the central line. The value of these areas may be found from statistical tables or in any particular case by integrating the formula for the curve between limits, e.g. the area of the tail of the curve beyond a point x_1 on one side of the curve is given by

$$\int_{x_1}^{+\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{x^2}{2\sigma^2}} dx$$

$$\text{or } \frac{1}{\sqrt{2\pi}\,\sigma} \int_{x_1}^{+\infty} e^{-\frac{x^2}{2\sigma^2}} dx$$

*This value of z is not to be confused with Fisher's z^1 which is the hyperbolic arc tangent of r the correlation coefficient, i.e. $z^1 = \tanh^{-1} r$. This transformation gives more reliable results than r under certain conditions.

APPENDIX IV

THE SPEARMAN RANKS FORMULA FOR CORRELATION

$$\rho = 1 - \frac{6\Sigma d^2}{N(N^2-1)}$$

If σ^2 is the square of the standard deviation of a set of n ranks

$$\sigma^2 = \frac{(1^2 + 2^2 + 3^2 + 4^2 + \ldots + n^2)}{n} - \left(\frac{1 + 2 + 3 + 4 + \ldots n}{n}\right)^2$$

i.e. $$\sigma^2 = \frac{\Sigma n^2}{n} - \left(\frac{\Sigma n}{n}\right)^2$$

By adding the identities

$$\begin{aligned}
(n+1)^3 - n^3 &= 3n^2 + 3n + 1 \\
n^3 - (n-1)^3 &= 3(n-1)^2 + 3(n-1) + 1 \\
(n-1)^3 - (n-2)^3 &= 3(n-2)^2 + 3(n-1) + 1 \\
&\cdots\cdots\cdots \\
2^3 - 1^3 &= 3\cdot 1^2 + 3\cdot 1 + 1 \\
(n+1)^3 - 1^3 &= 3\Sigma n^2 + 3\Sigma n + n
\end{aligned}$$

Σn is the sum of the first n natural numbers, i.e. half the sum of the first and last number multiplied by the number of terms.

$$\Sigma n = \frac{n(n+1)}{2}$$

Substituting in the identity $n^3 + 3n^2 + 3n = 3\Sigma n^2 + 3\Sigma n + n$

$$3\Sigma n^2 = n^3 + 3n^2 + 3n - 3\frac{(n+1)}{2} - n$$

$$6\Sigma n^2 = 2n^3 + 6n^2 + 6n - 3n(n+1) - 2n$$

$$6\Sigma n^2 = 2n^3 + 3n^2 + n$$

$$= n(2n+1)(n+1)$$

$$\Sigma n^2 = \frac{n(2n+1)(n+1)}{6}$$

Substituting in our variance formula

$$\sigma^2 = \frac{\Sigma n^2}{n} - \left(\frac{\Sigma n}{n}\right)^2 \quad \text{we have}$$

$$\sigma^2 = \frac{n(2n+1)(n+1)}{6n} - \frac{n^2(n+1)^2}{4n^2}$$

$$= \frac{n^2 - 1}{12}$$

Now if ρ is the correlation coefficient between pairs of scores assuming that the variabilities and the means of the two sets of ranks are equal

$$\rho = 1 - \frac{\Sigma d^2}{2n\sigma^2}$$

which gives

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

by substituting for σ^2

This can also be *demonstrated* in a simpler way:

It would appear from the following identities:

$$1^2 + 3^2 = \tfrac{1}{6} \times 4 \times (4^2 - 1)$$
$$2^2 + 4^2 = \tfrac{1}{6} \times 5 \times (5^2 - 1)$$
$$1^2 + 3^2 + 5^2 = \tfrac{1}{6} \times 6 \times (6^2 - 1)$$
$$2^2 + 4^2 + 6^2 = \tfrac{1}{6} \times 7 \times (7^2 - 1)$$

that the sum of the squares of consecutive odd numbers or consecutive even numbers beginning with 2 as far as N−1 is $\frac{1}{6}N(N^2-1)$.

Now consider the following cases of perfect negative rank correlation (i.e. $\rho = -1$):

Case (N is odd)	*Order of Merit* (rank) in subject P	*Order of Merit* (rank) in subject R	*Difference in rank squared* d^2
A	1	7	6^2
B	2	6	4^2
C	3	5	2^2
D	4	4	0^2
E	5	3	2^2
F	6	2	4^2
G	7	1	6^2

Case (N is even)	*Order of Merit* (rank) in subject P	*Order of Merit* (rank) in subject R	*Difference in Rank Squared* d^2
A	1	8	7^2
B	2	7	5^2
C	3	6	3^2
D	4	5	1^2
E	5	4	1^2
F	6	3	3^2
G	7	2	5^2
H	8	1	7^2

It will be seen that in both cases where there is perfect negative correlation $\Sigma d^2 = \frac{1}{3}N(N^2 - 1)$. Obviously when the ranks are identical and there is perfect positive correlation $\Sigma d^2 = 0$; therefore it is reasonable to suppose that when there is zero correlation (i.e. half way between -1 and $+1$) that Σd^2 is half way between $\frac{1}{3}N(N^2 - 1)$ and 0, i.e. $\frac{1}{6}N(N^2 - 1)$.

Now if Σd^2 were determined by chance alone (no correlation) it would have the value $\frac{1}{6}N(N^2 - 1)$.

Thus $\dfrac{\Sigma d^2}{\frac{1}{6}N(N^2 - 1)}$ gives a measure of the *lack* of association between the ranks or the variance of the set of ranks.

The correlation coefficient $\rho = 1 - \dfrac{\Sigma d^2}{\frac{1}{6}N(N^2 - 1)}$

which can be written $1 - \dfrac{6\Sigma d^2}{N(N^2 - 1)}$

APPENDIX V

A NOTE ON CORRELATION AND REGRESSION LINES

CONSIDER N numbers $A_1, A_2, A_3 \ldots$ Denote their mean by $\bar{A}$ and the differences or deviations of the numbers from their mean by $a_1, a_2, a_3 \ldots$ etc., so that the mean of these is 0.

The standard deviation is $\left(\frac{\Sigma a^2}{N}\right)^{\frac{1}{2}}$

If $\sigma_a = 1$ the numbers a_1, a_2, a_3 are said to be in standard measure. (Alternatively, we could have achieved the same result by dividing the deviations from the mean by the standard deviation.)

Consider a second set of N numbers $B_1, B_2, B_3 \ldots$ and in the same way derive from them $b_1, b_2, b_3 \ldots$ and σ_b the standard deviation of this set.

The coefficient of correlation $r_{ab} = \frac{\Sigma ab}{N\sigma_a \sigma_b}$ by definition.

Consider the identity

$$\begin{aligned}\Sigma\Sigma(a_p b_q - a_q b_p)^2 &= (\Sigma a^2)(\Sigma b^2) - (\Sigma ab)^2 \\ &= N^2\sigma_a{}^2\sigma_b{}^2 (1 - r_{ab}{}^2)\end{aligned}$$

It follows from this that if $r_{ab} = \pm 1$, $\frac{a_p}{b_p} = \frac{a_q}{a_q}$ for all values of p and q giving a straight line relationship between each A and the corresponding value of B. (Note that as the left-hand side of the identity, being a square, cannot be negative, r_{ab} cannot lie outside the limits -1 and $+1$.)

Normally no such exact linear relation exists but we may find the line of best fit by finding one which will make the sum of the squares of the distances of points from it a minimum.

Choose λ and μ so that $\Sigma(b - \lambda a - \mu)^2$ is a minimum.

Differentiating partially with respect to λ and μ, we obtain

$$-2\Sigma a(b - \lambda a - \mu) = 0 = -2\Sigma(b - \lambda a - \mu)$$

which as $\Sigma a = N\bar{a} - 0$ and similarly $\Sigma b = 0$

reduce to $-2(N\sigma_a\sigma_b r_{ab} - \lambda N\sigma_a^2) = 0 = -2(-N\mu)$

and thus $\lambda = \frac{\sigma_a r_{ab}}{\sigma_a}$ and $\mu = 0$

The line of regression of B on A is given by

$$b = \left(\frac{\sigma_b r_{ab}}{\sigma_a}\right) a$$

and the line of regression of A on B is given by

$$a = \left(\frac{\sigma_b r_{ab}}{\sigma_a}\right) b.$$

If the As and Bs are quite independent r_{ab} will approximate to zero if N is large enough. The converse is only true for a linear relationship. In the case of the parabolic curve $b^2 = a$, r_{ab} would equal 0 and we should use the correlation ratio instead of the coefficient. Thus, independence involves zero correlation, if N is large enough, but zero correlation does not necessarily imply independence.[1]

[1] Adapted from 'Mathematics and Psychology', Piaggio, *Mathematical Gazette*, February 1933 This paper also contains 'An analysis of the factor *g*, if it exists'

APPENDIX VI

AN EASY PROOF THAT THE COEFFICIENT OF CORRELATION IS LESS THAN UNITY

Suppose that x and y are the standardized scores of a person in two subjects respectively and that there are N persons taking the test.

From the process of standardization it follows that Σx^2 and Σy^2 both equal N.

The algebraic identity $(x-y)^2 = x^2 - 2xy + y^2$

may be written $2xy = x^2 + y^2 - (x-y)^2$

$$\therefore 2\Sigma xy = \Sigma x^2 + \Sigma y^2 - \Sigma(x-y)^2$$

$$2\Sigma xy = N + N - \Sigma(x-y)^2$$

$$\Sigma xy = N - \tfrac{1}{2}\Sigma(x-y)^2$$

$$\therefore \text{Correlation coefficient } \frac{\Sigma xy}{N} = 1 - \frac{\Sigma(x-y)^2}{2N}$$

But the square $(x-y)^2$ must always be positive

$\therefore \frac{\Sigma xy}{N}$ must always be less than 1 unless $x = y$ in which case the term $\frac{\Sigma(x-y)^2}{2N}$ vanishes and $\Sigma xy = N$.

BIBLIOGRAPHY

FOR an account of recent work in factorial analysis the student is recommended to read Professor Godfrey H. Thomson's *Factorial Analysis of Human Ability*, Second Edition. This admirably-written and impartial work not only gives a clear account of the ideas of various workers in this field in terms of fairly simple mathematics, but it does much to reconcile some of the apparently different ideas of the American authorities.

The Measurement of Abilities by P. E. Vernon is the best work extant on the statistics of mental testing, marking and the 'new' examining.

The Factors of the Mind by Sir Cyril Burt is an excellent work on the measurement of mental traits and it should be read in conjunction with Thomson's book which we have mentioned above.

The original research in educational matters which appears in *The British Journal of Educational Psychology* very often makes great use of statistical methods and in particular the analysis of variance, *in recent issues. A new section of The British Journal which will* be devoted to statistical matters solely is soon to make its appearance.

FAIRLY EASY WORKS

Mental Tests. Ballard. University of London Press.

Group Tests of Intelligence. Ballard. University of London Press.

The Science of Marking. Thomas. Murray.

Statistical Calculations for Beginners. Chambers. Cambridge University Press.

How to Calculate a Correlation. Thomson. Harrap.

A First Course in Statistics. Lindquist. Harrap.

The Distribution and Relations of Educational Abilities. Burt. King.

A Guide to Mental Testing. Cattell. University of London Press.

The Selection of Children for Secondary Education. Davies and Jones. Harrap.

Some Recent Work in Factorial Analysis and a Retrospect. Thomson. Harrap.

The Testing of Intelligence. Ed. Hamley. Evans.
An Introduction to the Computation of Statistics. Dawson. University of London Press.
Elementary Matrices. Turnbull and Aitken. Blackie.
Intelligence, Concrete and Abstract. Alexander. British Journal of Psychology Monograph.
An Examination of Examinations. Hartog and Rhodes. Macmillan.
Statistics in Psychology and Education. Garrett. Longmans-Green.
The Reliability of Examinations. Valentine and Emmett. University of London Press.
Essentials of Mental Measurement. Brown and Thomson. Cambridge University Press.
Research in Education. Oliver. Allen & Unwin.
Mental and Scholastic Tests. Burt. King.
Elementary Statistics. Levy and Preidel. Nelson.

MODERATELY DIFFICULT WORKS

The Measurement of Abilities.[1] Vernon. University of London Press.
The Factorial Analysis of Human Ability[1] (Second Edition). Thomson. University of London Press.
The Factors of the Mind[1] (Second Edition). Burt. University of London Press.
The Abilities of Man. Spearman. Macmillan.
An Introduction to the Theory of Statistics.[2] Yule & Kendall. Griffin.
Statistical Methods. Snedecor. Iowa College.
Statistical Method. Kelley. Macmillan.
Statistical Procedures and their Mathematical Bases.[2] Peters and Van Voorhis. McGraw-Hill.
Statistical Methods for Research Workers.[2] Fisher. Oliver & Boyd.
Design of Experiments.[2] Fisher. Oliver & Boyd.
Methods of Statistical Analysis. Goulden. Wiley.
The Vectors of Mind. Thurstone. University of Chicago Press.
Primary Mental Abilities. Thurstone. University of Chicago Press.
Psychometric Methods. Guilford. McGraw-Hill.

[1] The first three works are of great importance to students of education and psychology.
[2] These books contain useful sets of statistical tables.

Tables for Statisticians and Biometricians. Pearson. Cambridge.
The Methods of Statistics.[1] Tippett. Oxford.
Statistical Tables. Fisher & Yates. Oliver & Boyd.
Statistical Analysis in Educational Research. Lindquist. Harrap.
Statistical Methods Applied to Education. Rugg. Houghton.
Statistical Analysis in Biology. Mather. Methuen.

[1] This contains an excellent explanation of analysis of variance.

INDEX